

Lingue fossili: filogenesi linguistica computazionale, glottocronologia bayesiana e alberi evolutivi delle famiglie linguistiche

Le lingue come organismi evolutivi e il metodo comparativo

Il paradigma evolutivo nello studio delle lingue

La linguistica storica ha sviluppato nel corso degli ultimi due secoli un corpus metodologico straordinariamente raffinato per ricostruire la storia delle lingue e i rapporti genealogici che le legano. L'analogia tra l'evoluzione biologica e il cambiamento linguistico, già intuita dai primi comparatisti dell'Ottocento, ha trovato una formulazione sempre più rigorosa grazie all'applicazione di metodi quantitativi e computazionali che traggono ispirazione diretta dalla filogenesi molecolare. Le lingue, al pari delle specie viventi, si trasformano nel tempo, si differenziano in varietà distinte, conservano tracce delle loro origini comuni e possono essere rappresentate mediante alberi filogenetici che ne visualizzano le relazioni di discendenza.

Questa prospettiva evolutiva non è meramente metaforica: esistono profonde analogie strutturali tra i meccanismi di trasmissione e mutazione del materiale genetico e quelli della trasmissione linguistica. Entrambi i sistemi sono caratterizzati da ereditarietà verticale da una generazione all'altra, da innovazioni casuali che si accumulano nel tempo, da pressioni selettive che favoriscono alcune varianti a scapito di altre, e da tassi di cambiamento che possono essere modellati statisticamente. Naturalmente, le lingue presentano anche caratteristiche distintive che le differenziano dagli organismi biologici, in particolare la capacità di prestito orizzontale attraverso il contatto linguistico, fenomeno che complica notevolmente la ricostruzione filogenetica ma che può essere affrontato con strumenti analitici appropriati.

Gli alberi filogenetici delle grandi famiglie linguistiche

Le famiglie linguistiche costituiscono gruppi di lingue discendenti da un antenato comune, il cui grado di differenziazione riflette la profondità temporale della divergenza. Tre delle famiglie più studiate e meglio documentate sono l'indoeuropeo, l'austroasiatico e il niger-congo, ciascuna delle quali comprende più di cento lingue viventi distribuite su vasti territori geografici.

La famiglia indoeuropea rappresenta il caso più celebre e meglio compreso della linguistica storica. Essa include lingue parlate da oltre tre miliardi di persone, distribuite dall'Europa

occidentale all'India settentrionale, e comprende rami come il germanico, il romanzo, il celtico, lo slavo, il baltico, l'indo-iranico, il greco, l'armeno, l'albanese e i rami estinti anatolico e tocharico. La ricostruzione del protoindoeuropeo, la lingua ancestrale da cui tutte queste varietà discendono, costituisce uno dei maggiori successi della linguistica comparativa. Gli studiosi hanno potuto ricostruire con notevole dettaglio il sistema fonologico, la morfologia flessiva estremamente complessa e una porzione significativa del lessico di questa lingua parlata probabilmente tra il 4000 e il 6000 avanti Cristo, senza che di essa sia rimasta alcuna testimonianza scritta diretta.

La famiglia austronesiana, con oltre milleduecento lingue attestate, rappresenta la famiglia linguistica geograficamente più estesa del pianeta, coprendo un'area che va dal Madagascar nell'Oceano Indiano fino all'Isola di Pasqua nell'Oceano Pacifico orientale, passando per l'Indonesia, le Filippine, la Malesia, il Taiwan e la Polinesia. Questa straordinaria diffusione geografica è il risultato di una delle più spettacolari espansioni umane della preistoria, avvenuta attraverso sofisticate tecniche di navigazione oceanica. Il protoaustronesiano è stato ricostruito con successo e le relazioni interne alla famiglia sono relativamente ben comprese, con il riconoscimento del ruolo centrale delle lingue formosane (Taiwan) come gruppo più conservativo e basale dell'intera famiglia.

La famiglia niger-congo, che include oltre millecinquecento lingue parlate nell'Africa subsahariana, presenta una complessità interna considerevole e rappresenta forse il caso più problematico tra le grandi famiglie linguistiche. Il sottogruppo bantu, con oltre cinquecento lingue, è relativamente ben studiato e mostra chiari segni di espansione relativamente recente dal bacino del fiume Niger verso l'Africa orientale e meridionale, probabilmente associata alla diffusione dell'agricoltura e della metallurgia del ferro. Tuttavia, i rapporti filogenetici tra i diversi rami della famiglia niger-congo, e persino l'effettiva esistenza di questa famiglia come unità genealogica coerente, rimangono oggetto di dibattito tra gli specialisti.

Il metodo comparativo e le corrispondenze sistematiche

Il metodo comparativo rappresenta il fondamento metodologico della linguistica storica e si basa sull'identificazione di corrispondenze regolari tra i suoni delle lingue imparentate. Il principio fondamentale è che i cambiamenti fonetici operano in modo sistematico: quando un suono si trasforma in un altro in una determinata posizione fonetica, questa trasformazione coinvolge tutte le occorrenze di quel suono in quella posizione, producendo serie di corrispondenze regolari nel lessico delle lingue figlie.

La legge di Grimm, formulata da Jacob Grimm nel 1822, rappresenta il primo e più celebre esempio di legge fonetica regolare. Essa descrive una serie sistematica di mutazioni consonantiche che hanno caratterizzato il passaggio dal protoindoeuropeo alle lingue germaniche. In particolare, le occlusive sorde del protoindoeuropeo divennero fricative sorde nel protogermanico: l'occlusiva bilabiale sorda *p* divenne la fricativa labiodentale *f*, l'occlusiva dentale sorda *t* divenne la fricativa dentale *th* (come nell'inglese "thin"), e l'occlusiva velare sorda *k* divenne la fricativa velare *h*. Parallelamente, le occlusive sonore del protoindoeuropeo divennero occlusive sonore nel protogermanico: *b* divenne *p*, *d* divenne *t*, e *g* divenne *k*. Infine, le occlusive sonore aspirate del protoindoeuropeo persero l'aspirazione divenendo semplici occlusive sonore nel protogermanico.

Queste trasformazioni possono essere verificate sistematicamente attraverso serie di cognati. Consideriamo il protoindoeuropeo *pāter che designava il padre: in latino divenne "pater", in greco "patēr", in sanscrito "pitā", tutte lingue che conservarono la p iniziale protoindoeuropea. Nel germanico, invece, la p iniziale divenne f secondo la legge di Grimm, producendo il gotico "fadar", l'antico inglese "fæder", l'antico alto tedesco "fater", da cui il tedesco moderno "Vater" e l'inglese moderno "father". La corrispondenza p : f tra il latino e le lingue germaniche non è limitata a questo esempio ma si manifesta in centinaia di cognati: latino "piscis" corrisponde all'inglese "fish", latino "pes" (piede) all'inglese "foot", latino "plenus" all'inglese "full", e così via.

La legge di Verner, formulata da Karl Verner nel 1875, rappresentò un perfezionamento cruciale della legge di Grimm, risolvendo apparenti irregolarità che avevano tormentato i linguisti per decenni. Verner dimostrò che le fricative sorde prodotte dalla legge di Grimm diventavano sonore quando la sillaba precedente non portava l'accento nel protoindoeuropeo, condizione che poteva essere dedotta dal sanscrito vedico. Questa scoperta rafforzò enormemente il principio della regolarità assoluta dei cambiamenti fonetici, principio che divenne il dogma centrale della scuola neogrammatica e che rimane fondamentale nella linguistica storica contemporanea.

L'identificazione dei cognati e la base lessicale della comparazione

I cognati sono parole in lingue diverse che discendono da una forma ancestrale comune attraverso regolare trasmissione ereditaria. L'identificazione dei cognati costituisce il primo passo essenziale nella ricostruzione filogenetica, ma richiede grande cautela metodologica per evitare falsi positivi dovuti a somiglianze casuali o a prestiti.

La serie di cognati per "padre" nelle lingue indoeuropee illustra perfettamente il concetto: inglese "father", tedesco "Vater", olandese "vader", gotico "fadar", latino "pater", greco "patēr", sanscrito "pitā", persiano "pedar", irlandese "athair", armeno "hayr" sono tutti riflessi regolari del protoindoeuropeo *pāter. Le differenze fonetiche tra queste forme non sono casuali ma riflettono le leggi fonetiche specifiche di ciascun ramo della famiglia, come la legge di Grimm per il germanico, l'indebolimento e perdita della p iniziale nel celtico, la palatalizzazione in armeno.

Un esempio ancora più trasparente è fornito dalla parola per "madre": inglese "mother", tedesco "Mutter", olandese "moeder", latino "māter", greco "mētēr", sanscrito "mātā", persiano "mādar", irlandese "máthair", lituano "motė", russo "mat'", tutte forme che risalgono al protoindoeuropeo *mātēr. La straordinaria somiglianza fonetica di queste forme attraverso lingue separate da migliaia di chilometri e millenni di evoluzione indipendente costituisce una prova schiacciante della loro origine comune.

È fondamentale distinguere i cognati dai prestiti linguistici. Il termine italiano "computer", per esempio, è presente in moltissime lingue del mondo con forma pressoché identica, ma non costituisce un cognato bensì un prestito recente dall'inglese. I cognati devono mostrare corrispondenze fonetiche regolari secondo le leggi fonetiche note, mentre i prestiti spesso mantengono la forma della lingua donatrice anche quando questa viola le corrispondenze attese. Inoltre, i prestiti tendono a coinvolgere maggiormente elementi del lessico culturale e tecnologico, mentre il lessico di base tende a essere più resistente al prestito.

La lessicostastica e le liste Swadesh

La lessicostastica è un approccio quantitativo alla classificazione linguistica che si basa sul calcolo della percentuale di cognati condivisi nel lessico di base. Morris Swadesh, negli anni Cinquanta del Novecento, sviluppò liste standardizzate di concetti fondamentali che si presumono essere universalmente lessicalizzati e relativamente resistenti al prestito. La lista Swadesh originale comprendeva duecento elementi, successivamente ridotti a cento nella versione più comunemente utilizzata.

Questi elementi includono pronomi personali ("io", "tu", "noi"), numerali bassi ("uno", "due"), termini per parti del corpo ("occhio", "mano", "bocca", "dente"), fenomeni naturali ("sole", "luna", "acqua", "fuoco"), azioni fondamentali ("mangiare", "bere", "dormire", "vedere", "morire"), proprietà di base ("grande", "piccolo", "lungo", "rosso", "bianco") e altri concetti elementari che si ritiene facciano parte del nucleo lessicale di ogni lingua.

L'idea sottostante è che questo lessico di base, essendo semanticamente elementare e pragmaticamente indispensabile, venga trasmesso prevalentemente per via ereditaria piuttosto che per prestito, e quindi la percentuale di cognati condivisi in questa lista rifletta in modo relativamente diretto il grado di parentela genealogica tra le lingue. Due lingue strettamente imparentate, come l'italiano e lo spagnolo, mostreranno una percentuale molto alta di cognati condivisi nella lista Swadesh, tipicamente superiore all'ottanta per cento. Lingue dello stesso ramo ma più distanti, come l'italiano e il rumeno, mostreranno percentuali più basse ma ancora significative. Lingue di rami diversi della stessa famiglia, come l'italiano e il russo, mostreranno percentuali ancora più ridotte ma comunque rilevabili. Lingue di famiglie diverse non dovrebbero mostrare cognati genuini, e qualsiasi somiglianza osservata sarebbe attribuibile a coincidenza o a prestiti antichi.

Nella pratica, l'applicazione della lessicostastica richiede competenza specialistica per identificare correttamente i cognati e distinguerli dai prestiti, dalle convergenze casuali e dalle derivazioni da radici diverse che producono forme superficialmente simili. Per esempio, l'italiano "molto" e il francese "beaucoup" esprimono lo stesso concetto ma non sono cognati: il primo deriva dal latino "multum", il secondo da "bellus" e "coup". Questo tipo di divergenza semantica e convergenza formale fortuita può complicare notevolmente l'analisi lessicostastica.

Nonostante queste difficoltà metodologiche, la lessicostastica ha fornito strumenti utili per una prima classificazione quantitativa delle lingue e per identificare possibili relazioni genealogiche che meritano indagine più approfondita mediante il metodo comparativo tradizionale. Essa ha inoltre posto le basi per lo sviluppo della glottocronologia, un tentativo più ambizioso di datare la divergenza linguistica sulla base della percentuale di cognati conservati.

Glottocronologia e le sue critiche

Il modello glottocronologico classico

La glottocronologia rappresenta un tentativo audace di trasformare la linguistica storica da disciplina puramente qualitativa a scienza quantitativa capace di datare la divergenza tra le

lingue con la stessa precisione con cui i metodi radiometrici datano i reperti archeologici. Il modello fu sviluppato principalmente da Morris Swadesh negli anni Cinquanta, basandosi sull'assunzione che il lessico di base si deteriori a un tasso costante nel tempo, analogamente al decadimento radioattivo.

L'equazione fondamentale della glottocronologia è espressa come $t = (\log C) / (2 \log r)$, dove t rappresenta il tempo trascorso dalla divergenza espresso in millenni, C rappresenta la percentuale di cognati condivisi tra due lingue nel lessico di base, e r rappresenta il tasso di ritenzione per millennio, cioè la frazione del lessico che sopravvive dopo mille anni.

Swadesh, sulla base dello studio di lingue per le quali esisteva documentazione storica che permetteva di calibrare il modello, propose un valore di ritenzione r pari a circa 0,86 o 0,81 a seconda che si utilizzasse la lista di cento o duecento elementi. Questo significa che, in media, dopo mille anni una lingua avrebbe conservato circa l'ottantuno-ottantasei per cento del suo lessico di base originale, mentre il rimanente quattordici-diciannove per cento sarebbe stato sostituito da innovazioni.

Il fattore due al denominatore della formula deriva dal fatto che stiamo confrontando due lingue che si sono evolute indipendentemente da un antenato comune, quindi il deterioramento lessicale è avvenuto in entrambe le linee di discendenza. Se vogliamo calcolare quanto tempo è trascorso dalla divergenza, dobbiamo considerare che entrambe le lingue hanno subito cambiamenti lessicali, e la percentuale di cognati condivisi riflette il prodotto dei tassi di conservazione nelle due linee separate.

Consideriamo un esempio concreto: supponiamo di confrontare due lingue romanze come l'italiano e il francese sulla lista Swadesh di cento elementi e di trovare che condividono settantacinque cognati, quindi $C = 0,75$. Utilizzando la formula con $r = 0,86$, otteniamo $t = (\log 0,75) / (2 \log 0,86) = (-0,125) / (2 \times -0,066) = (-0,125) / (-0,132) = 0,95$ millenni, cioè circa novecentocinquanta anni. Poiché sappiamo dalle fonti storiche che il latino volgare si frammentò nelle diverse lingue romanze tra il sesto e il nono secolo dell'era comune, una datazione intorno al mille-millecinquecento anni fa sarebbe ragionevolmente accurata, suggerendo che in questo caso specifico il metodo glottocronologico fornisce una stima ragionevole.

Tuttavia, se applicassimo lo stesso metodo a lingue più distanti, per esempio l'italiano e il russo (entrambe indoeuropee ma di rami diversi), e trovassimo che condividono circa venti cognati su cento, quindi $C = 0,20$, otterremmo $t = (\log 0,20) / (2 \log 0,86) = (-0,699) / (-0,132) = 5,3$ millenni, suggerendo una divergenza circa cinquemilatrecento anni fa. Questo risultato è problematico perché la separazione tra il ramo italico e quello slavo risale probabilmente a un periodo molto più antico, forse settemila-ottomila anni fa o anche oltre, indicando che il modello sottostima sistematicamente la profondità temporale quando le percentuali di cognati sono basse.

Le critiche fondamentali al modello glottocronologico

La glottocronologia è stata oggetto di critiche devastanti sin dalla sua formulazione iniziale, e oggi è considerata dalla maggior parte dei linguisti storici come uno strumento metodologicamente difettoso che produce datazioni inaffidabili. Le critiche si concentrano

principalmente su tre aspetti: l'assunzione di tassi di cambiamento costanti, i problemi di calibrazione, e la difficoltà di identificare oggettivamente i cognati.

La critica più fondamentale riguarda l'assunzione di un tasso di cambiamento costante. Contrariamente all'analogia con il decadimento radioattivo, che è un processo fisico governato da leggi probabilistiche invarianti, il cambiamento linguistico è influenzato da innumerevoli fattori storico-sociali che variano enormemente tra le diverse lingue e i diversi periodi. Lingue con sistemi di scrittura ben sviluppati e tradizioni letterarie prestigiose tendono a cambiare più lentamente del previsto dal modello, perché la norma scritta esercita un effetto conservatore sulla lingua parlata. L'islandese moderno, per esempio, è rimasto straordinariamente simile al norreno medievale grazie al prestigio della letteratura delle saghe e all'isolamento geografico dell'Islanda, mentre il norvegese parlato nella Norvegia continentale si è trasformato molto più rapidamente.

Al contrario, lingue parlate da comunità piccole e geograficamente frammentate, o lingue sottoposte a intense pressioni da parte di lingue dominanti, possono cambiare molto più rapidamente. Le lingue creole, che emergono da situazioni di contatto estremo, possono svilupparsi nell'arco di una o due generazioni, un tasso di innovazione lessicale che renderebbe completamente inapplicabile il modello glottocronologico standard. Anche all'interno della stessa famiglia linguistica, i tassi di cambiamento possono variare di diversi ordini di grandezza: alcune lingue romanze come il sardo hanno conservato caratteristiche molto arcaiche del latino volgare, mentre altre come il francese hanno subito trasformazioni fonologiche e morfologiche radicali.

Studi empirici su lingue con documentazione storica estesa hanno dimostrato che i tassi di sostituzione lessicale variano tipicamente di un fattore compreso tra due e cinque anche all'interno della stessa famiglia linguistica. Questo significa che l'errore introdotto dall'assunzione di un tasso costante può facilmente portare a datazioni errate del cento-duecento per cento o più, rendendo le stime glottocronologiche quasi inutili per scopi scientifici rigorosi.

Un secondo problema cruciale riguarda la calibrazione del tasso di ritenzione. Swadesh basò le sue stime su un numero limitato di lingue per le quali esisteva documentazione storica, principalmente lingue indoeuropee e alcune lingue amerindie. Non è affatto chiaro che il tasso calibrato su queste lingue sia universalmente applicabile a tutte le famiglie linguistiche del mondo. Inoltre, anche per le lingue utilizzate per la calibrazione esistono ambiguità significative nella datazione precisa della divergenza: quando esattamente il latino volgare si frammentò nelle diverse lingue romanze? Si tratta di un processo graduale che si estese per diversi secoli, non di un evento puntuale, rendendo problematica qualsiasi calibrazione precisa.

Un terzo problema, più tecnico ma non meno importante, riguarda l'identificazione dei cognati. In teoria, la lista Swadesh dovrebbe contenere solo lessico di base resistente al prestito, ma nella pratica le lingue in intenso contatto si prestano parole anche dal lessico fondamentale. L'inglese, per esempio, ha sostituito molti termini germanici originali con prestiti dal francese normanno dopo la conquista del 1066, inclusi alcuni elementi che sarebbero sulla lista Swadesh. Se non si riconosce la natura di prestito di queste forme e le si contano erroneamente come cognati condivisi con le altre lingue romanze, si

sovrastimerebbe artificialmente la relazione tra l'inglese e le lingue romanze, portando a conclusioni genealogiche completamente errate.

Inoltre, esiste il problema delle convergenze casuali. Quando si confrontano lingue molto distanti temporalmente, la probabilità che forme foneticamente simili emergano per pura coincidenza aumenta significativamente. Distinguere i cognati genuini dalle somiglianze casuali richiede l'applicazione rigorosa del metodo comparativo e la verifica di corrispondenze fonetiche sistematiche, ma questo livello di analisi dettagliata spesso non è possibile quando si lavora con lingue scarsamente documentate o con ricostruzioni ipotetiche di protolinguaggi.

Tentativi di raffinamento e l'eredità della glottocronologia

Nonostante queste critiche devastanti, alcuni ricercatori hanno tentato di raffinare il metodo glottocronologico piuttosto che abbandonarlo completamente. Sergei Starostin e altri linguisti della scuola moscovita hanno proposto versioni modificate che tentano di tenere conto della variazione nei tassi di cambiamento tra diverse categorie semantiche: alcuni concetti, come i pronomi personali o i numerali bassi, sembrano essere estremamente stabili attraverso il tempo, mentre altri sono più soggetti a sostituzione. Assegnando pesi diversi a categorie diverse, si spera di ottenere stime più accurate.

Altri approcci hanno tentato di modellare esplicitamente la variazione nei tassi di cambiamento piuttosto che assumere un tasso costante. Questi modelli a "orologio rilassato" (relaxed clock), mutuati dalla filogenesi molecolare, permettono ai tassi di cambiamento di variare tra i diversi rami dell'albero evolutivo, e stimano questi tassi simultaneamente con la topologia dell'albero e le datazioni dei nodi. Questo approccio, implementato in software come BEAST, rappresenta un miglioramento significativo rispetto alla glottocronologia classica ed è diventato lo standard metodologico nella filogenesi linguistica bayesiana contemporanea.

L'eredità più importante della glottocronologia, al di là delle sue specifiche assunzioni metodologiche difettose, è stata quella di stimolare il pensiero quantitativo nella linguistica storica e di porre il problema della datazione assoluta delle divergenze linguistiche come una questione scientifica legittima che richiede metodi rigorosi. Anche se il modello originale di Swadesh è stato in gran parte abbandonato, l'impulso verso approcci quantitativi e computazionali ha portato allo sviluppo di metodi filogenetici molto più sofisticati che rappresentano oggi la frontiera della ricerca nella linguistica storica.

Filogenesi bayesiana e metodi computazionali avanzati

I fondamenti della filogenesi bayesiana applicata alle lingue

La filogenesi bayesiana rappresenta un salto qualitativo rispetto ai metodi tradizionali di ricostruzione filogenetica, introducendo un quadro statistico rigoroso per l'inferenza di alberi evolutivi e la datazione dei nodi. L'approccio bayesiano si distingue per la sua capacità di incorporare esplicitamente l'incertezza nelle stime, di combinare diverse fonti di informazione

attraverso distribuzioni a priori, e di esplorare lo spazio delle possibili topologie filogenetiche in modo sistematico.

Il teorema di Bayes costituisce il fondamento matematico del metodo. In termini semplici, esso permette di calcolare la probabilità di un'ipotesi (in questo caso, un particolare albero filogenetico con specifiche lunghezze dei rami e datazioni dei nodi) dato un insieme di dati osservati (per esempio, le forme lessicali nelle lingue moderne e alcuni punti di calibrazione storica). La formula fondamentale è $P(\theta|D) = [P(D|\theta) \times P(\theta)] / P(D)$, dove θ rappresenta i parametri del modello (topologia dell'albero, lunghezze dei rami, tassi di sostituzione, età del nodo radice), D rappresenta i dati osservati, $P(\theta|D)$ è la distribuzione a posteriori che vogliamo calcolare, $P(D|\theta)$ è la verosimiglianza dei dati dato il modello, $P(\theta)$ è la distribuzione a priori che riflette le nostre conoscenze precedenti, e $P(D)$ è una costante di normalizzazione.

Nella pratica, la distribuzione a posteriori non può essere calcolata analiticamente per problemi complessi come la ricostruzione filogenetica, dove lo spazio dei parametri è enormemente vasto. Si ricorre quindi a metodi di campionamento come il Markov Chain Monte Carlo (MCMC), che esplora lo spazio dei parametri in modo stocastico, visitando le regioni di alta probabilità a posteriori con frequenza proporzionale alla loro probabilità. Dopo un periodo di "burn-in" durante il quale la catena converge verso la distribuzione stazionaria, i campioni estratti forniscono un'approssimazione empirica della distribuzione a posteriori, permettendo di calcolare stime puntuali (come la media a posteriori o il massimo a posteriori) e intervalli di credibilità che quantificano l'incertezza.

Il software BEAST e l'implementazione pratica

BEAST (Bayesian Evolutionary Analysis Sampling Trees) è diventato lo strumento standard per l'analisi filogenetica bayesiana in linguistica, dopo essere stato originariamente sviluppato per la filogenesi molecolare. Il software implementa modelli estremamente flessibili che permettono di specificare distribuzioni a priori su tutti i parametri di interesse, di incorporare punti di calibrazione temporale basati su evidenze storiche o archeologiche, e di utilizzare modelli a orologio rilassato che non assumono tassi di cambiamento costanti.

Un'analisi tipica in BEAST richiede la specifica di diversi componenti cruciali. Innanzitutto, è necessario definire il modello di sostituzione dei caratteri, che descrive come le forme lessicali si trasformano nel tempo. Per dati lessicali binari (presenza o assenza di un cognato), si utilizza tipicamente un modello stocastico semplice di nascita e morte dei cognati. Per dati più complessi, come sequenze di caratteri fonetici, si possono utilizzare modelli più sofisticati analoghi ai modelli di sostituzione nucleotidica utilizzati in biologia molecolare.

In secondo luogo, è necessario specificare un modello a priori per la topologia dell'albero. I modelli più comunemente utilizzati sono il modello di nascita pura (Yule process), che assume un tasso costante di speciazione linguistica, e il modello nascita-morte, che incorpora anche un tasso di estinzione delle lingue. Questi modelli generano distribuzioni a priori sulle possibili topologie e sulle lunghezze dei rami che riflettono assunzioni biologicamente (o linguisticamente) plausibili sui processi di diversificazione.

In terzo luogo, è fondamentale specificare il modello di orologio molecolare. L'orologio rilassato lognormale è particolarmente popolare in linguistica: esso assume che i tassi di sostituzione lungo i diversi rami dell'albero siano distribuiti secondo una distribuzione lognormale, permettendo una variazione sostanziale nei tassi pur mantenendo una struttura probabilistica trattabile. Questo è cruciale perché, come discusso precedentemente, i tassi di cambiamento linguistico variano enormemente tra le diverse lingue.

Infine, l'analisi richiede la specificazione di punti di calibrazione temporale che ancorino l'albero a una scala temporale assoluta. Questi punti di calibrazione sono tipicamente basati su attestazioni storiche documentate: per esempio, sappiamo dalla documentazione storica che il latino si frammentò nelle diverse lingue romanze approssimativamente tra il 400 e il 700 dell'era comune, quindi possiamo utilizzare questa informazione per calibrare il nodo che rappresenta il protolatino volgare. Similmente, possiamo utilizzare le prime attestazioni scritte di una lingua come limite inferiore per la sua esistenza.

La calibrazione mediante eventi storici documentati

La calibrazione costituisce uno degli aspetti più delicati e potenzialmente problematici della filogenesi bayesiana. La qualità delle datazioni ottenute dipende crucialmente dalla qualità e dall'appropriatezza dei punti di calibrazione utilizzati. Nel caso delle lingue indoeuropee, esistono diversi punti di calibrazione ragionevolmente robusti basati su documentazione storica.

La frammentazione del latino nelle lingue romanze è probabilmente il punto di calibrazione più utilizzato. Le ultime iscrizioni in latino classico e le prime testimonianze di forme volgari nettamente differenziate geograficamente suggeriscono che il processo di frammentazione fu sostanzialmente completato entro il settimo-ottavo secolo. Tuttavia, questo processo non fu istantaneo: la variazione dialettale esisteva già durante l'Impero Romano, e la transizione da un continuum dialettale latino a lingue romanze distinte fu graduale. Nella pratica, i ricercatori utilizzano tipicamente una distribuzione a priori sulla datazione di questo nodo piuttosto che un punto fisso, per esempio una distribuzione uniforme tra il 400 e il 700 dell'era comune, o una distribuzione normale centrata sul 500 con una deviazione standard appropriata.

Un altro punto di calibrazione importante riguarda la separazione tra il ramo indo-iranico e gli altri rami indoeuropei. Le più antiche attestazioni del sanscrito vedico e dell'avestico iranico risalgono approssimativamente al 1500-1000 avanti Cristo, fornendo un limite superiore sicuro per l'età di queste lingue. La separazione tra il ramo indiano e quello iranico deve essere anteriore, e considerazioni archeologiche relative alla cultura andronovo delle steppe euroasiatiche suggeriscono una datazione intorno al 2000-2500 avanti Cristo.

La separazione del ramo anatolico (ittita, luvio) dal resto della famiglia indoeuropea è particolarmente importante perché il ramo anatolico mostra caratteristiche molto arcaiche e potrebbe essere il primo ramo a essersi separato dal ceppo principale. Le tavolette ittite di Kültepe in Anatolia forniscono testimonianze scritte che risalgono al 1900-1700 avanti Cristo, e la presenza di prestiti indoeuropei in sumerico e accadico suggerisce che parlanti di lingue indoeuropee erano già presenti nell'Anatolia orientale nel terzo millennio avanti Cristo.

Tuttavia, l'uso di attestazioni scritte come calibrazioni comporta difficoltà intrinseche. La scrittura cattura tipicamente uno stadio relativamente conservatore e standardizzato della lingua, che può essere significativamente diverso dalle varietà parlate contemporanee. Inoltre, l'assenza di attestazioni scritte non implica l'assenza della lingua: molte lingue indoeuropee furono parlate per secoli o millenni prima di essere trascritte. Quindi le prime attestazioni scritte forniscono solo limiti inferiori piuttosto laxi sull'età delle lingue, introducendo incertezza sostanziale nelle calibrazioni.

Gli studi pionieristici: Gray e Atkinson (2003)

Il lavoro seminale di Russell Gray e Quentin Atkinson, pubblicato su *Nature* nel 2003, rappresentò la prima applicazione sistematica della filogenesi bayesiana computerizzata allo studio delle lingue indoeuropee, producendo risultati che riaccesero un dibattito vecchio di decenni sull'origine geografica e temporale della famiglia.

Gray e Atkinson compilarono un dataset di cognati per ottantasette lingue indoeuropee (più tre lingue utilizzate come outgroup) basato su liste lessicali standardizzate. Codificarono i cognati come caratteri binari: per ogni significato nella lista Swadesh e per ogni cognato identificato, assegnarono il valore 1 alle lingue che possedevano quel cognato e 0 alle altre lingue. Questo produsse una matrice di migliaia di caratteri binari che poteva essere analizzata con metodi filogenetici.

Utilizzando il software bayesiano con un modello a orologio rilassato e calibrazioni basate su attestazioni storiche note, Gray e Atkinson ottennero una datazione per l'età del protoindoeuropeo di circa ottomila-novemilacinquecento anni prima del presente, significativamente più antica delle stime tradizionali che collocavano il protoindoeuropeo intorno a seimila anni fa. Inoltre, la loro analisi filogeografica, che incorporava informazioni sulla distribuzione geografica delle lingue, suggeriva un'origine del protoindoeuropeo nell'Anatolia orientale, supportando la cosiddetta "ipotesi anatolica" proposta dall'archeologo Colin Renfrew.

Secondo questa ipotesi, la diffusione delle lingue indoeuropee sarebbe stata associata all'espansione dell'agricoltura dall'Anatolia verso l'Europa e l'Asia nel Neolitico, intorno al 7000-6000 avanti Cristo. Questo contrastava nettamente con l'"ipotesi delle steppe", sostenuta dalla maggior parte dei linguisti storici e degli archeologi, secondo la quale il protoindoeuropeo sarebbe stato parlato nelle steppe ponto-caspiche intorno al 4000-3000 avanti Cristo, e la sua diffusione sarebbe stata associata all'addomesticamento del cavallo e all'invenzione del carro.

Il lavoro di Gray e Atkinson generò un dibattito acceso e numerose critiche. Alcuni linguisti misero in dubbio la qualità del dataset di cognati, segnalando possibili errori di codifica e inclusioni problematiche. Altri criticarono le assunzioni del modello bayesiano e l'interpretazione delle calibrazioni. Tuttavia, il lavoro ebbe il merito fondamentale di introdurre metodi quantitativi rigorosi in un campo tradizionalmente dominato da approcci qualitativi, e di dimostrare che la linguistica storica poteva beneficiare degli stessi strumenti computazionali che avevano rivoluzionato la biologia evolutiva.

L'analisi di Bouckaert et al. (2012): filogeografia e diffusione spaziale

Un lavoro successivo ancora più ambizioso fu pubblicato da Remco Bouckaert e collaboratori su Science nel 2012. Questo studio espanse notevolmente il dataset precedente, includendo centotré lingue indoeuropee e fornendo un'analisi filogeografica molto più sofisticata che modellava esplicitamente la diffusione spaziale delle lingue nel tempo.

Il modello filogeografico utilizzato da Bouckaert e colleghi tratta la posizione geografica come un carattere che evolve lungo l'albero filogenetico secondo un processo di diffusione spaziale. Ogni lingua è associata a una località geografica (tipicamente il centroide dell'area in cui è o era parlata), e il modello stima la probabile posizione geografica degli antenati non osservati a ciascun nodo dell'albero. Questo permette di ricostruire non solo le relazioni genealogiche tra le lingue, ma anche le rotte migratorie e i pattern di espansione geografica che hanno portato alla distribuzione attuale.

L'analisi produsse un supporto statistico molto forte per l'ipotesi anatolica, con una probabilità a posteriori superiore al novantacinque per cento per un'origine del protoindoeuropeo in Anatolia. La datazione stimata per il nodo radice era coerente con quella di Gray e Atkinson, collocandosi intorno agli ottomila-novemila anni prima del presente. Le ricostruzioni filogeografiche mostravano pattern di espansione compatibili con la diffusione dell'agricoltura neolitica, con rami linguistici che si irradiavano dall'Anatolia verso l'Europa occidentale, la penisola balcanica e, più tardi, verso le steppe orientali.

Questo lavoro fu accolto con grande interesse ma anche con critiche sostanziali. Don Ringe, uno dei massimi esperti di linguistica storica indoeuropea, pubblicò una risposta dettagliata in cui argomentava che le datazioni molto antiche erano artefatti di problemi metodologici, in particolare dell'inadeguatezza delle calibrazioni utilizzate. David Anthony, archeologo specializzato nelle culture delle steppe, sostenne che l'evidenza archeologica e genetica supportava molto più fortemente l'ipotesi delle steppe rispetto a quella anatolica.

Studi genetici pubblicati successivamente, in particolare le analisi del DNA antico delle popolazioni delle steppe, hanno in effetti fornito supporto molto forte all'ipotesi delle steppe, mostrando che le popolazioni europee moderne derivano in parte significativa da migrazioni dalle steppe ponto-caspiche iniziate intorno al 3000 avanti Cristo, esattamente come previsto dall'ipotesi delle steppe. Questo ha portato molti ricercatori a concludere che, nonostante la sofisticazione dei metodi bayesiani, le datazioni molto antiche prodotte dalle analisi filogenetiche riflettono probabilmente problemi nei dati o nelle assunzioni del modello piuttosto che la realtà storica.

Contatto linguistico, reti evolutive e filogenesi strutturale

Il problema del contatto linguistico e del trasferimento orizzontale

Una delle differenze più fondamentali tra l'evoluzione biologica e l'evoluzione linguistica riguarda il meccanismo di trasferimento orizzontale, noto in linguistica come prestito o contatto linguistico. Mentre gli organismi trasmettono il loro materiale genetico prevalentemente in modo verticale da genitori a figli all'interno della stessa specie, le lingue

possono scambiare elementi lessicali, fonologici, morfologici e sintattici attraverso il contatto tra comunità linguistiche diverse. Questo fenomeno complica enormemente la ricostruzione filogenetica, perché le somiglianze osservate tra le lingue possono riflettere non solo eredità comune da un antenato condiviso, ma anche prestiti dovuti a contatto.

Il prestito lessicale è il tipo più comune e più facilmente identificabile di trasferimento orizzontale. Le lingue europee moderne contengono migliaia di prestiti reciproci: l'inglese ha preso massicciamente dal francese normanno dopo la conquista del 1066, il russo ha incorporato numerosi termini dal francese durante il diciottesimo e diciannovesimo secolo, le lingue balcaniche (greco, albanese, rumeno, bulgaro) condividono un substrato lessicale comune dovuto a secoli di contatto nella penisola balcanica. In casi estremi, il prestito può coinvolgere anche il lessico fondamentale che normalmente si assume essere resistente al contatto: l'inglese "sky" (cielo) è un prestito dal norreno, sostituendo il termine germanico originale che sopravvive nel tedesco "Himmel".

Il problema metodologico cruciale è che i prestiti possono facilmente essere scambiati per cognati se non si applica il metodo comparativo con estremo rigore. Due lingue geograficamente vicine ma genealogicamente distanti possono condividere molti termini lessicali a causa del contatto, creando un segnale filogenetico spurio che può ingannare gli algoritmi di ricostruzione filogenetica. Per esempio, le lingue baltiche e finniche condividono numerosi termini a causa di millenni di contatto, ma appartengono a famiglie completamente diverse (indoeuropea e uralica rispettivamente). Se questi prestiti non vengono identificati correttamente e rimossi dal dataset, l'analisi filogenetica potrebbe erroneamente suggerire una relazione genealogica stretta.

Il contatto linguistico non si limita al lessico ma può coinvolgere anche la fonologia, la morfologia e persino la sintassi. Alcuni linguisti parlano di "aree linguistiche" (Sprachbund) per descrivere regioni geografiche in cui lingue genealogicamente non imparentate hanno sviluppato caratteristiche strutturali comuni attraverso il contatto prolungato. L'area linguistica balcanica è l'esempio più celebre: lingue come il greco, l'albanese, il rumeno, il bulgaro e il macedone, che appartengono a rami diversi della famiglia indoeuropea, condividono una serie di caratteristiche grammaticali insolite (come il futuro analitico formato con il verbo "volere", l'articolo determinativo posposto al nome, la perdita dell'infinito) che si sono diffuse per contatto.

Questo fenomeno pone una sfida fondamentale al modello ad albero della diversificazione linguistica. Un albero filogenetico assume che le lingue divergano per separazione da antenati comuni e che ogni innovazione si trasmetta verticalmente lungo i rami dell'albero. Il contatto linguistico viola questa assunzione, introducendo somiglianze che attraversano i confini dei rami e creando pattern reticolari piuttosto che strettamente arborescenti.

I modelli a rete e le rappresentazioni non arborescenti

Per affrontare il problema del contatto linguistico, sono stati sviluppati metodi che rappresentano le relazioni tra le lingue mediante reti piuttosto che alberi puri. Una rete filogenetica è un grafo che può contenere non solo nodi di biforcazione (che rappresentano la separazione di una lingua ancestrale in due o più lingue figlie) ma anche eventi di fusione o trasferimento laterale che rappresentano il contatto linguistico.

Il metodo NeighborNet è uno degli approcci più utilizzati per generare reti filogenetiche dalle distanze lessicali tra le lingue. Questo metodo produce grafici a separazione (splits graphs) che visualizzano segnali filogenetici potenzialmente conflittuali. In un grafo NeighborNet, i pattern ad albero puro producono ramificazioni nette, mentre i pattern reticolari producono parallelogrammi o strutture a rete che indicano segnali conflittuali nei dati, tipicamente causati da prestiti o da incertezza nella ricostruzione filogenetica.

Per esempio, un'analisi NeighborNet delle lingue indoeuropee mostra tipicamente una struttura prevalentemente arborescente per i principali rami della famiglia, confermando che il modello ad albero è sostanzialmente appropriato a questo livello di analisi. Tuttavia, all'interno di alcune regioni geografiche come i Balcani o il subcontinente indiano, dove il contatto linguistico è stato particolarmente intenso, emergono strutture reticolari che riflettono l'influenza del contatto.

Un altro approccio consiste nei modelli di mistura o ammixture, mutuati dalla genetica delle popolazioni. Questi modelli rappresentano ciascuna lingua come una combinazione ponderata di sorgenti ancestrali. Formalmente, si può scrivere la lingua L come $L = \alpha \times S1 + (1-\alpha) \times S2$, dove $S1$ e $S2$ sono lingue sorgente e α è il coefficiente di mistura che indica la proporzione di materiale linguistico derivato dalla prima sorgente rispetto alla seconda. Questi modelli sono particolarmente appropriati per descrivere lingue creole o pidgin, che emergono dalla fusione di due o più lingue in situazioni di contatto estremo, ma possono essere applicati anche a situazioni di contatto meno drammatico.

L'applicazione pratica di questi modelli richiede algoritmi sofisticati per l'inferenza dei parametri. Tipicamente si utilizza un approccio di massima verosimiglianza o bayesiano, cercando i valori dei coefficienti di mistura che massimizzano la probabilità dei dati osservati o che hanno la più alta probabilità a posteriori. L'identificazione di eventi di ammixture può fornire informazioni preziose sulla storia sociale delle comunità linguistiche, rivelando episodi di contatto intenso che potrebbero non essere evidenti dalla documentazione storica.

La filogenesi strutturale e il World Atlas of Language Structures

Un approccio complementare alla filogenesi lessicale consiste nell'utilizzare caratteristiche strutturali o tipologiche piuttosto che il lessico per ricostruire le relazioni tra le lingue. Il World Atlas of Language Structures (WALS) è una risorsa fondamentale in questo ambito, fornendo dati sistematici su migliaia di caratteristiche strutturali per 2679 lingue del mondo.

Le caratteristiche codificate nel WALS includono aspetti della fonologia (per esempio, la presenza o assenza di determinati fonemi, la complessità del sistema consonantico o vocalico), della morfologia (per esempio, il tipo di marcatura del caso, la presenza di genere grammaticale, i sistemi di numero), della sintassi (l'ordine delle parole, la posizione degli aggettivi rispetto al nome, la struttura delle frasi relative), e della semantica lessicale (per esempio, i sistemi di colore, i sistemi di parentela).

La logica sottostante alla filogenesi strutturale è che queste caratteristiche grammaticali profonde sono meno soggette a prestito rispetto al lessico, e quindi dovrebbero fornire un segnale filogenetico più puro delle relazioni genealogiche. Effettivamente, mentre è relativamente facile prendere in prestito parole da un'altra lingua, è molto più difficile

adottare caratteristiche grammaticali fondamentali: richiede tipicamente un bilinguismo diffuso e prolungato nella comunità.

Tuttavia, l'applicazione della filogenesi ai dati strutturali presenta difficoltà metodologiche specifiche. Innanzitutto, il numero di caratteristiche tipologiche codificabili è molto più limitato rispetto al numero di cognati potenzialmente identificabili: mentre un dataset lessicale può facilmente includere migliaia di cognati, i dataset tipologici tipicamente contengono solo centinaia di caratteristiche. Questo riduce il potere statistico dell'analisi.

In secondo luogo, le caratteristiche strutturali non sono necessariamente indipendenti: molte di esse sono correlate per ragioni funzionali o tipologiche. Per esempio, le lingue con ordine delle parole verbo-oggetto tendono ad avere preposizioni piuttosto che posposizioni, e ad avere l'aggettivo dopo il nome piuttosto che prima. Queste correlazioni violano l'assunzione di indipendenza dei caratteri che sta alla base della maggior parte dei metodi filogenetici, potenzialmente distorto le ricostruzioni.

In terzo luogo, le caratteristiche strutturali sembrano essere più omoplastiche del lessico, cioè tendono a evolversi convergentemente in modo indipendente in lingue non imparentate con frequenza relativamente alta. Questo è particolarmente vero per caratteristiche fonologiche: per esempio, i toni si sono evoluti indipendentemente in numerose famiglie linguistiche completamente non imparentate (sinotibetano, niger-congo, austriaco, amerindio), probabilmente perché esistono pressioni funzionali che favoriscono lo sviluppo di sistemi tonali in determinate condizioni.

Studi empirici che hanno confrontato alberi filogenetici basati sul lessico con alberi basati su caratteristiche strutturali hanno prodotto risultati contrastanti. In alcuni casi, i due tipi di dati producono topologie filogenetiche sostanzialmente concordanti, suggerendo che entrambi riflettono la storia genealogica reale delle lingue. In altri casi, tuttavia, emergono discrepanze significative, che possono riflettere tassi diversi di evoluzione per i due tipi di caratteri, oppure l'influenza del contatto linguistico o dell'adattamento funzionale.

Metodi computazionali: approcci basati su distanza e su caratteri

I metodi computazionali per la ricostruzione filogenetica si dividono tradizionalmente in due grandi categorie: metodi basati su distanza e metodi basati su caratteri. Ciascun approccio ha vantaggi e svantaggi specifici, e la scelta del metodo appropriato dipende dalla natura dei dati e dagli obiettivi dell'analisi.

I metodi basati su distanza partono da una matrice di distanze tra tutte le coppie di lingue nel dataset. La distanza può essere definita in vari modi: per dati lessicali, si utilizza comunemente la proporzione di cognati non condivisi, mentre per caratteristiche strutturali si può utilizzare la distanza di Hamming (proporzione di caratteri differenti). Data questa matrice di distanze, l'algoritmo cerca di costruire un albero che rappresenti al meglio le relazioni di distanza osservate.

Il metodo UPGMA (Unweighted Pair Group Method with Arithmetic Mean) è uno degli algoritmi più semplici di questa categoria. Esso procede gerarchicamente, unendo iterativamente le coppie di lingue o gruppi di lingue più simili, e calcolando le distanze ai nuovi nodi come media aritmetica delle distanze ai nodi figlie. Il metodo assume

implicitamente un orologio molecolare stretto, cioè tassi di evoluzione costanti in tutti i rami, il che lo rende inappropriato per molti dataset linguistici dove questa assunzione è violata.

Il metodo Neighbor-Joining (NJ) è più sofisticato e non assume un orologio molecolare. Esso cerca di minimizzare la lunghezza totale dell'albero, selezionando a ogni passo la coppia di nodi da unire in modo da ridurre al massimo questa lunghezza. Il metodo NJ è computazionalmente molto efficiente e produce risultati ragionevoli in molti casi, ma ha lo svantaggio di essere un metodo "greedy" che fa scelte localmente ottimali senza garantire l'ottimalità globale della soluzione.

I metodi basati su caratteri, al contrario, lavorano direttamente sui caratteri discreti osservati nelle lingue piuttosto che su distanze aggregate. Questi metodi costruiscono esplicitamente un modello dell'evoluzione dei caratteri e cercano l'albero che meglio spiega i pattern osservati secondo un determinato criterio di ottimalità.

Il metodo della parsimonia massima cerca l'albero che richiede il minor numero totale di cambiamenti evolutivi (mutazioni) per spiegare i dati osservati. L'idea è che, a parità di altre condizioni, l'ipotesi più semplice che spiega i dati è da preferire. Questo approccio ha il vantaggio di essere concettualmente semplice e di non richiedere assunzioni dettagliate sul processo evolutivo, ma ha lo svantaggio di poter essere statisticamente inconsistente in certe condizioni (per esempio, in presenza di "long-branch attraction", dove rami lunghi dell'albero tendono a essere erroneamente raggruppati).

I metodi di massima verosimiglianza sono statisticamente più rigorosi. Essi richiedono la specificazione esplicita di un modello probabilistico dell'evoluzione dei caratteri, che include parametri come i tassi di transizione tra stati diversi del carattere. Dato questo modello, si calcola la verosimiglianza di ciascun albero candidato, cioè la probabilità di osservare i dati attuali supponendo che quell'albero sia vero, e si seleziona l'albero con la verosimiglianza massima. Questo approccio ha basi statistiche solide ma richiede assunzioni modellistiche che possono essere difficili da giustificare e da verificare empiricamente.

Validazione, prospettive future e sintesi metodologica

La validazione attraverso famiglie linguistiche note

Una delle critiche più acute rivolte ai metodi filogenetici computazionali in linguistica riguarda la difficoltà di validare le ricostruzioni prodotte. A differenza della biologia molecolare, dove spesso si può confrontare la filogenesi molecolare con il registro fossile o con altre linee di evidenza indipendenti, in linguistica storica raramente disponiamo di una "verità di riferimento" indipendente contro cui testare le nostre ricostruzioni. Tuttavia, esiste un'importante eccezione: le famiglie linguistiche per le quali possediamo documentazione storica estesa che ci permette di conoscere con certezza le relazioni filogenetiche.

La famiglia romanza rappresenta il caso ideale per questo tipo di validazione. Conosciamo con certezza che l'italiano, lo spagnolo, il francese, il portoghese, il rumeno e le altre lingue romanze discendono tutte dal latino, e possediamo una documentazione scritta quasi continua che copre l'intero processo di diversificazione dal latino volgare alle lingue romanze moderne. Inoltre, conosciamo con ragionevole precisione le datazioni relative e assolute dei

principali eventi di diversificazione: la separazione tra il ramo orientale (balcano-romanzo) e occidentale, la formazione dei principali sottogruppi (ibero-romanzo, gallo-romanzo, italo-romanzo), le diverse ondate di innovazioni che hanno attraversato diverse parti del territorio romanzo.

Diversi studi hanno utilizzato questo "esperimento naturale" per testare l'accuratezza dei metodi filogenetici computazionali. Il procedimento standard consiste nel compilare un dataset di cognati o di caratteristiche strutturali per le lingue romanze, applicare i vari algoritmi di ricostruzione filogenetica (distanza-based, parsimonia, massima verosimiglianza, bayesiano), e confrontare la topologia dell'albero ricostruito con la topologia vera nota dalla documentazione storica. Si può anche confrontare le datazioni stimate per i nodi dell'albero con le datazioni storiche note.

I risultati di questi studi di validazione sono stati generalmente incoraggianti ma hanno anche rivelato limitazioni importanti. Per quanto riguarda la topologia, la maggior parte dei metodi ricostruisce correttamente le relazioni di base tra le principali lingue romanze, confermando per esempio che il francese e lo spagnolo sono più strettamente imparentati tra loro di quanto ciascuno di essi lo sia con il rumeno. Tuttavia, emergono spesso errori nei dettagli: alcune lingue possono essere posizionate nel ramo sbagliato, o la struttura interna di alcuni sottogruppi può essere ricostruita incorrettamente.

Le datazioni assolute si sono rivelate più problematiche. Come discusso precedentemente, i metodi glottocronologici classici tendono a produrre datazioni con margini di errore molto ampi, e la direzione dell'errore può variare: in alcuni casi le datazioni sono sistematicamente troppo recenti, in altri troppo antiche. I metodi bayesiani con orologio rilassato producono risultati generalmente più accurati, ma l'accuratezza dipende crucialmente dalla qualità delle calibrazioni utilizzate e dalle assunzioni del modello a priori.

Un risultato importante emerso da questi studi di validazione è che dataset diversi possono produrre alberi significativamente diversi. Gli alberi basati sul lessico di base tendono a essere più congruenti con le relazioni genealogiche note rispetto agli alberi basati su caratteristiche fonetiche o morfologiche, che sembrano essere più influenzati da convergenze areali dovute al contatto. Questo suggerisce che, come già affermato dai linguisti storici tradizionali, il lessico di base fornisce effettivamente il segnale filogenetico più affidabile, anche se non è immune da problemi.

La cross-validazione e i test di robustezza

Oltre alla validazione su famiglie note, esistono tecniche statistiche generali di validazione che possono essere applicate anche quando non disponiamo di una topologia di riferimento certa. La cross-validazione è una tecnica standard in statistica e machine learning che consiste nel dividere il dataset in un insieme di addestramento (training set) e un insieme di test, costruire il modello sull'insieme di addestramento, e valutarne le prestazioni predittive sull'insieme di test.

Nel contesto della filogenesi linguistica, la cross-validazione può essere implementata rimuovendo una porzione dei dati (per esempio, un sottoinsieme dei cognati o delle lingue), ricostruendo l'albero filogenetico sui dati rimanenti, e poi verificando quanto bene questo albero predice i dati rimossi. Se il metodo è robusto e i dati contengono un segnale

filogenetico chiaro, l'albero ricostruito dovrebbe avere buon potere predittivo anche sui dati non utilizzati per la sua costruzione.

Un'altra tecnica importante è il bootstrap, una procedura di ricampionamento che permette di quantificare l'incertezza statistica nella ricostruzione filogenetica. Il bootstrap procede generando molti dataset artificiali dello stesso tamanho del dataset originale campionando con reimmissione dai caratteri originali. Per ciascuno di questi dataset bootstrap si ricostruisce un albero filogenetico, e alla fine si calcola la frequenza con cui ciascun raggruppamento (clade) appare negli alberi bootstrap. Questa frequenza, chiamata supporto bootstrap, fornisce una misura della robustezza statistica di quel particolare raggruppamento: valori alti (tipicamente sopra il settanta-ottanta per cento) indicano che il raggruppamento è fortemente supportato dai dati, mentre valori bassi indicano che c'è incertezza significativa.

Nel framework bayesiano, l'incertezza è quantificata naturalmente attraverso le probabilità a posteriori. Dopo che la catena MCMC ha campionato migliaia o milioni di alberi dalla distribuzione a posteriori, si può calcolare la frequenza con cui ciascun raggruppamento appare in questi campioni. Questa probabilità a posteriori fornisce una misura bayesiana del supporto per quel raggruppamento, ed è generalmente considerata più accurata statisticamente rispetto al supporto bootstrap, anche se richiede la specificazione di distribuzioni a priori che possono influenzare i risultati.

I test di ipotesi filogenetiche

Un aspetto cruciale della scienza è la capacità di testare ipotesi alternative in modo rigoroso. Nella filogenesi linguistica, spesso ci troviamo di fronte a ipotesi concorrenti sulle relazioni tra le lingue o sulla datazione di eventi cruciali. I metodi statistici moderni permettono di confrontare formalmente queste ipotesi alternative.

Nel framework della massima verosimiglianza, si possono utilizzare test di rapporto di verosimiglianza (likelihood ratio tests) per confrontare modelli annidati. Per esempio, se vogliamo testare se un particolare raggruppamento di lingue è supportato dai dati, possiamo confrontare la verosimiglianza massima dell'albero migliore che include quel raggruppamento con la verosimiglianza massima dell'albero migliore che non lo include. Se la differenza di verosimiglianza è sufficientemente grande, possiamo rifiutare l'ipotesi che il raggruppamento non esista.

Nel framework bayesiano, il confronto tra modelli alternativi si effettua tipicamente attraverso i fattori di Bayes, che quantificano quanto i dati favoriscono un modello rispetto a un altro. Il fattore di Bayes è il rapporto tra le verosimiglianze marginali (evidence) dei due modelli, dove la verosimiglianza marginale è ottenuta integrando la verosimiglianza su tutti i possibili valori dei parametri pesati dalle loro probabilità a priori. Valori del fattore di Bayes molto maggiori di uno indicano supporto forte per il primo modello, valori molto minori di uno indicano supporto per il secondo modello, e valori vicini a uno indicano che i dati non permettono di distinguere tra i due modelli.

Questi test statistici formali rappresentano un progresso importante rispetto agli approcci tradizionali della linguistica storica, che spesso si basavano su giudizi qualitativi e su dibattiti retorici piuttosto che su confronti quantitativi rigorosi. Tuttavia, è importante riconoscere che

anche i test statistici più sofisticati sono validi solo nella misura in cui le assunzioni sottostanti sono soddisfatte, e che risultati statisticamente significativi non equivalono necessariamente a conclusioni linguisticamente sensate.

Integrazione con evidenze archeologiche e genetiche

Una delle tendenze più promettenti nella ricerca contemporanea sulla preistoria linguistica è l'integrazione di evidenze provenienti da discipline diverse: linguistica, archeologia, genetica, paleoclimatologia. L'idea è che ciascuna di queste discipline fornisca una finestra parziale sulla storia umana, e che la convergenza di evidenze indipendenti possa portare a ricostruzioni storiche più robuste.

L'archeologia fornisce informazioni sulla cultura materiale, sui pattern di insediamento, sulle rotte commerciali, sulla diffusione di tecnologie e pratiche agricole. Questi dati possono essere correlati con le ricostruzioni linguistiche per testare ipotesi sulle espansioni di popolazioni che parlavano determinate famiglie linguistiche. Per esempio, l'espansione delle lingue bantu nell'Africa subsahariana è correlata archeologicamente con la diffusione della metallurgia del ferro e di specifiche tecniche agricole, fornendo un contesto storico-culturale per l'espansione linguistica.

La genetica delle popolazioni, e in particolare gli studi sul DNA antico estratto da resti scheletrici, ha rivoluzionato la nostra comprensione delle migrazioni preistoriche. Le analisi del DNA antico hanno dimostrato che l'Europa moderna è il risultato della fusione di almeno tre componenti ancestrali principali: cacciatori-raccoglitori mesolitici indigeni, agricoltori neolitici provenienti dall'Anatolia e dal Vicino Oriente, e pastori delle steppe dell'età del bronzo. Questa terza componente, associata alla cultura Yamnaya delle steppe ponto-caspiche, porta una firma genetica che è massima nelle popolazioni dell'Europa settentrionale e si riduce gradualmente verso sud, un pattern geografico che correla bene con la distribuzione delle lingue indoeuropee.

Studi genetici recenti hanno anche permesso di datare questi eventi di mescolanza genetica con precisione crescente. Le analisi mostrano che la componente genetica delle steppe appare improvvisamente in Europa centrale e settentrionale intorno al 3000-2500 avanti Cristo, esattamente nel periodo in cui l'ipotesi delle steppe colloca l'espansione dei parlanti indoeuropei. Questo fornisce un supporto indipendente molto forte all'ipotesi delle steppe e pone serie difficoltà per l'ipotesi anatolica, che prevedrebbe un flusso genetico molto più antico associato alla diffusione neolitica dell'agricoltura.

L'integrazione di queste diverse linee di evidenza richiede cautela metodologica. Correlazione non implica causalità, e pattern simili osservati in domini diversi non provano necessariamente che i processi sottostanti siano collegati. Tuttavia, quando evidenze indipendenti convergono sistematicamente verso lo stesso quadro storico, la robustezza delle conclusioni è notevolmente rafforzata.

Prospettive future e direzioni di ricerca

Il campo della filogenesi linguistica computazionale è in rapida evoluzione, con numerosi sviluppi metodologici e concettuali in corso. Alcune delle direzioni più promettenti per la ricerca futura includono:

L'incorporazione di modelli più realistici del cambiamento linguistico. I modelli attuali, mutuati dalla biologia molecolare, trattano i cambiamenti come eventi stocastici indipendenti, ma sappiamo che il cambiamento linguistico è spesso strutturato e correlato. Per esempio, certi cambiamenti fonetici tendono a verificarsi in cascata (chain shifts), e l'adozione di una innovazione morfologica può rendere più probabile l'adozione di altre innovazioni correlate. Modelli che incorporano queste dipendenze potrebbero produrre ricostruzioni più accurate.

Lo sviluppo di metodi specificamente progettati per gestire il contatto linguistico. Mentre i metodi a rete attuali rappresentano un passo nella direzione giusta, tendono a essere descrittivi piuttosto che esplicitamente modellati. Servirebbero modelli generativi che rappresentino esplicitamente i processi di prestito e di convergenza areale, e che permettano di stimare quantitativamente l'importanza relativa dell'eredità verticale rispetto al trasferimento orizzontale nella storia di ciascuna lingua.

L'utilizzo di dataset più ricchi e più accurati. Nonostante decenni di lavoro, i dataset esistenti di cognati e di caratteristiche strutturali contengono ancora molti errori e lacune. Progetti di digitalizzazione su larga scala, come quello che ha prodotto il database Lexibank, stanno rendendo disponibili dataset di qualità molto superiore che coprono centinaia di famiglie linguistiche. L'applicazione di metodi computazionali a questi dati migliorati dovrebbe produrre ricostruzioni più robuste.

L'integrazione esplicita di informazioni fonologiche dettagliate. La maggior parte degli studi filogenetici attuali codifica i cognati semplicemente come presenza/assenza, ignorando le informazioni ricche contenute nelle corrispondenze fonetiche sistematiche. Metodi che incorporano modelli espliciti dell'evoluzione fonologica, basati sui principi della fonetica articolatoria e acustica, potrebbero estrarre molto più segnale filogenetico dai dati lessicali.

Lo sviluppo di approcci di machine learning e intelligenza artificiale. Tecniche moderne come le reti neurali profonde hanno dimostrato capacità straordinarie nell'identificare pattern complessi in dati ad alta dimensionalità. Mentre queste tecniche sono state utilizzate finora principalmente per compiti sincronici in linguistica computazionale (come il riconoscimento vocale o la traduzione automatica), potrebbero essere adattate anche a problemi diacronici come l'identificazione automatica di cognati o la ricostruzione di protoforme.

Sintesi metodologica e raccomandazioni pratiche

Dopo questa disamina estesa dei metodi e delle problematiche della filogenesi linguistica computazionale, è utile fornire una sintesi delle best practices metodologiche che emergono dalla ricerca recente.

Per quanto riguarda la raccolta dei dati, è fondamentale utilizzare liste lessicali standardizzate come quella di Swadesh, ma integrandole con cognati aggiuntivi quando disponibili per aumentare il potere statistico. L'identificazione dei cognati deve essere effettuata da esperti qualificati utilizzando il metodo comparativo rigoroso, verificando l'esistenza di corrispondenze fonetiche sistematiche e escludendo i prestiti evidenti. È

preferibile essere conservativi, includendo solo cognati chiaramente identificabili, piuttosto che rischiare di includere falsi positivi che introducono rumore nel dataset.

Per l'analisi filogenetica, i metodi bayesiani con orologio rilassato implementati in software come BEAST rappresentano attualmente lo stato dell'arte, offrendo il miglior equilibrio tra rigore statistico e realismo biologico. Tuttavia, è essenziale testare la sensibilità dei risultati a diverse scelte di prior e di modello, e utilizzare misure di supporto statistico (probabilità a posteriori, supporto bootstrap) per quantificare l'incertezza.

Le calibrazioni temporali devono basarsi su evidenze storiche solide e devono essere specificate come distribuzioni probabilistiche piuttosto che come punti fissi, per riflettere realisticamente l'incertezza nelle datazioni storiche. È preferibile utilizzare multiple calibrazioni indipendenti quando possibile, e testare come i risultati cambiano quando si utilizzano calibrazioni diverse.

I risultati delle analisi filogenetiche computazionali devono sempre essere interpretati nel contesto più ampio delle conoscenze linguistiche, archeologiche e storiche. Un albero filogenetico statisticamente ben supportato che contraddice radicalmente ciò che sappiamo dalla ricostruzione comparativa tradizionale o dall'evidenza archeologica dovrebbe essere considerato sospetto e merita indagine attenta per identificare possibili problemi nei dati o nelle assunzioni del modello.

Infine, è importante riconoscere che tutti i metodi attuali hanno limitazioni significative, e che nessun singolo approccio fornisce risposte definitive. La triangolazione attraverso metodi multipli, dataset multipli, e discipline multiple offre le maggiori probabilità di arrivare a conclusioni robuste sulla preistoria linguistica umana. La linguistica storica computazionale è un campo giovane e in rapida evoluzione, e possiamo aspettarci progressi significativi nei prossimi anni man mano che vengono sviluppati metodi più sofisticati e dataset più ampi e accurati diventano disponibili.

Le lingue fossili come finestre sulla preistoria umana

Le lingue possono essere giustamente definite "fossili" nel senso che conservano tracce della loro storia evolutiva che possono essere ricostruite attraverso metodi scientifici rigorosi, analogamente a come i paleontologi ricostruiscono l'evoluzione delle forme di vita estinte dallo studio dei fossili. Il progresso della filogenesi linguistica computazionale negli ultimi due decenni ha trasformato quello che era un campo dominato da metodi qualitativi e dibattiti speculativi in una disciplina quantitativa con strumenti statistici sofisticati e protocolli metodologici rigorosi.

Tuttavia, è importante mantenere un appropriato senso di umiltà epistemologica. Le ricostruzioni filogenetiche, per quanto matematicamente sofisticate, sono basate su assunzioni e su dati che contengono inevitabilmente errori e incertezze. La storia linguistica reale è stata enormemente complessa, coinvolgendo non solo ramificazioni ad albero ma anche contatti, prestiti, convergenze, sostituzioni di lingue, e innumerevoli altri fenomeni che i nostri modelli attuali catturano solo imperfettamente.

Ciononostante, quando applicati con cura e interpretati con cautela, i metodi computazionali possono fornire intuizioni preziose sulla preistoria umana che sarebbero impossibili da

ottenere con altri mezzi. Le lingue che parliamo oggi sono il risultato di migliaia di anni di evoluzione culturale, e portano in sé le tracce delle migrazioni, dei contatti, delle separazioni e delle fusioni delle popolazioni che le hanno parlate. Decifrare queste tracce attraverso la filogenesi computazionale ci permette di ricostruire aspetti della storia umana che non hanno lasciato tracce archeologiche o scritte, illuminando i movimenti e le trasformazioni delle popolazioni umane nei millenni che hanno preceduto l'invenzione della scrittura.

La sfida continua è quella di perfezionare i nostri metodi, ampliare i nostri dataset, e integrare le nostre ricostruzioni linguistiche con le evidenze convergenti dall'archeologia, dalla genetica e dalle altre scienze storiche, costruendo una comprensione sempre più ricca e accurata della straordinaria odissea evolutiva che ha portato alla diversità linguistica che osserviamo nel mondo contemporaneo.

Applicazioni avanzate e casi di studio specifici

La ricostruzione del protoindoeuropeo: metodologia e risultati

La ricostruzione del protoindoeuropeo rappresenta forse il trionfo più celebrato della linguistica storica comparativa, e costituisce un banco di prova ideale per i metodi computazionali. Questa lingua ancestrale, parlata probabilmente tra il quinto e il terzo millennio avanti Cristo, non ha lasciato alcuna testimonianza scritta diretta, eppure i linguisti sono stati capaci di ricostruire con notevole dettaglio il suo sistema fonologico, la sua morfologia flessiva complessa e una porzione sostanziale del suo lessico.

Il sistema consonantico del protoindoeuropeo è stato ricostruito con particolare precisione. Esso includeva tre serie di occlusive: sorde (p , t , k , k^w), sonore (b , d , g , g^w), e sonore aspirate (b^h , d^h , g^h , g^{wh}), distinte in quattro punti di articolazione (labiale, dentale, velare semplice, velare labializzata). Questa ricostruzione si basa su corrispondenze sistematiche osservate attraverso tutte le lingue figlie: per esempio, il protoindoeuropeo $*b^her-$ "portare" produsse il sanscrito "bharati", il greco "pherō", il latino "ferō", il gotico "bairan", mostrando la regolare trasformazione della sonora aspirata in fricativa sonora in greco, in fricativa sorda in latino, in occlusiva sonora semplice nel germanico.

Il sistema vocalico ricostruito era relativamente semplice, con probabilmente solo cinque vocali di base ($*e$, $*o$, $*a$, $*i$, $*u$) più le loro controparti lunghe. Tuttavia, la scoperta della teoria laringale da parte di Ferdinand de Saussure nel 1879, confermata drammaticamente dalla decifrazione dell'ittita nel 1917, rivelò che il protoindoeuropeo possedeva anche consonanti laringali ($*h_1$, $*h_2$, $*h_3$) che coloravano e allungavano le vocali adiacenti prima di scomparire nella maggior parte delle lingue figlie. L'ittita conservò riflessi diretti di queste laringali come consonanti h , fornendo una conferma spettacolare di una predizione teorica fatta decenni prima sulla base di considerazioni puramente distribuzionali.

La morfologia verbale del protoindoeuropeo era straordinariamente complessa, con un sistema di coniugazione che distingueva almeno sei tempi o aspetti (presente, imperfetto, aoristo, perfetto, piuccheperfetto, futuro), tre modi (indicativo, congiuntivo, ottativo), tre voci (attiva, media, passiva), tre numeri (singolare, duale, plurale), e tre persone. Inoltre, il verbo mostrava ablaut, un sistema di alternanze vocaliche grammaticalizzate nella radice che codificava distinzioni aspettuali e derivazionali: la radice $*wed-$ "condurre (una sposa)"

poteva apparire come *wed- (grado pieno), *wod- (grado-o), *ud- (grado zero), producendo forme come il greco "edōka" (aoristo), "dedōka" (perfetto).

La morfologia nominale era altrettanto elaborata, con un sistema di declinazione che distingueva otto casi (nominativo, accusativo, genitivo, dativo, ablativo, locativo, strumentale, vocativo), tre numeri e tre generi (maschile, femminile, neutro). Il sistema di concordanza richiedeva che aggettivi, participi e pronomi concordassero con il nome in genere, numero e caso, producendo paradigmi flessivi di enorme complessità.

I metodi computazionali moderni hanno permesso di testare quantitativamente la coerenza interna di queste ricostruzioni. Algoritmi di massima verosimiglianza applicati a dataset di corrispondenze fonetiche possono stimare i tassi relativi di cambiamento per diversi tipi di suoni, confermando per esempio che le consonanti sibilanti e liquide tendono a essere più stabili delle occlusive, o che le vocali in sillaba accentata sono più conservate di quelle in sillaba atona. Questi pattern emergenti dai dati computazionali rafforzano e rendono più precise le intuizioni dei comparatisti tradizionali.

L'espansione austronesiana e la navigazione oceanica preistorica

La famiglia austronesiana offre un esempio affascinante di come la filogenesi linguistica possa illuminare eventi preistorici drammatici. L'espansione austronesiana, che portò popoli parlanti queste lingue dal Taiwan attraverso l'intero Pacifico fino al Madagascar e all'Isola di Pasqua, rappresenta una delle più spettacolari imprese di navigazione e colonizzazione della preistoria umana.

Gli studi filogenetici computazionali, combinati con evidenze archeologiche e genetiche, hanno permesso di ricostruire con notevole dettaglio la cronologia e le rotte di questa espansione. Le analisi bayesiane di Gray, Drummond e Greenhill (2009) su un dataset di quattrocento lingue austronesiane hanno stimato che il protoaustronesiano fu parlato nel Taiwan circa cinquemilaseicento anni fa. Da questa patria originaria, i popoli austronesiani si espansero verso sud nelle Filippine circa quattromilaquattrocento anni fa, quindi verso l'Indonesia e la Melanesia occidentale circa quattromila anni fa.

La fase più spettacolare fu l'espansione nel Pacifico remoto, iniziata circa tremilaquattrocento anni fa con la colonizzazione delle isole di Vanuatu, Nuova Caledonia e Fiji, associate archeologicamente alla cultura Lapita caratterizzata da ceramiche distintive con decorazioni elaborate. Da questo punto di partenza melanesiano, i navigatori polinesiani colonizzarono progressivamente l'intero triangolo polinesiano: le Samoa e le Tonga circa tremila anni fa, le Marchesi e la Società circa duemila anni fa, le Hawaii circa milleseicento anni fa, la Nuova Zelanda circa ottocento anni fa, e finalmente l'Isola di Pasqua, il punto più remoto, circa settecentocinquanta anni fa.

Questa cronologia filogenetica correla straordinariamente bene con le datazioni radiocarboniche dei primi insediamenti umani in queste isole, fornendo un esempio notevole di convergenza tra evidenze linguistiche, archeologiche e genetiche. Gli studi sul DNA mitocondriale e sui cromosomi Y delle popolazioni polinesiane moderne mostrano pattern di diversità genetica che riflettono fedelmente i pattern di diversificazione linguistica, con le popolazioni più orientali (hawaiani, maori, rapanui) che mostrano ridotta diversità genetica coerente con ripetuti effetti di fondazione durante la colonizzazione progressiva.

L'analisi filogenetica ha anche permesso di ricostruire aspetti del lessico del protoaustronesiano che illuminano la cultura materiale e le conoscenze tecnologiche di questi popoli. La ricostruzione di termini come *qabaŋ "barca a bilanciere", *layaR "vela", *tali "corda", *sakay "imbarcarsi", insieme a termini per piante coltivate (*pajay "riso", *qubi "yam", *niuR "cocco") e animali domestici (*manuk "pollo", *babuy "maiale"), dipinge un quadro di società marinare agricole con sofisticate tecnologie navali che permisero loro di colonizzare isole remote separate da centinaia o migliaia di chilometri di oceano aperto.

Il Niger-Congo e le controversie sulla profondità temporale

La famiglia Niger-Congo presenta sfide metodologiche particolarmente acute per la filogenesi computazionale, e il dibattito sulla sua validità come unità genealogica illustra bene i limiti dei metodi attuali quando applicati a lingue con grande profondità temporale e documentazione limitata.

La classificazione tradizionale di Joseph Greenberg, che raggruppava la maggior parte delle lingue dell'Africa subsahariana in una vasta famiglia Niger-Congo, è stata accettata da molti linguisti ma rimane controversa. Le lingue bantu, parlate in gran parte dell'Africa orientale e meridionale, formano chiaramente un sottogruppo coerente con solo tremila-quattromilacinquecento anni di differenziazione interna, come dimostrato sia dall'analisi comparativa tradizionale sia dagli studi filogenetici computazionali. Tuttavia, i rapporti tra il gruppo bantu e altri presunti rami del Niger-Congo, come l'atlantico, il mande, il gur, il kwa, sono molto meno chiari.

Il problema fondamentale è la profondità temporale. Se il protoNiger-Congo fu effettivamente parlato, probabilmente risale a ottomila-diecimila anni fa o anche oltre, un intervallo temporale paragonabile o superiore a quello che separa le lingue indoeuropee dal loro antenato comune. A queste profondità temporali, il segnale filogenetico nel lessico diventa estremamente debole: secoli di cambiamento fonologico hanno trasformato le parole cognate fino a renderle irriconoscibili, e la maggior parte del lessico originale è stata sostituita da innovazioni.

Gli studi filogenetici computazionali sul Niger-Congo hanno prodotto risultati contrastanti e statisticamente deboli. Mentre alcune analisi suggeriscono che esiste effettivamente un segnale filogenetico coerente che supporta l'unità della famiglia, altre concludono che i dati sono compatibili con multiple famiglie non imparentate che condividono somiglianze superficiali dovute a contatto, convergenza o coincidenza. I valori di supporto bootstrap o di probabilità a posteriori per molti dei raggruppamenti proposti sono bassi, indicando alta incertezza statistica.

Questo caso illustra un punto metodologico cruciale: i metodi filogenetici computazionali, per quanto sofisticati, non possono estrarre segnale da rumore quando il segnale è stato quasi completamente eroso dal tempo. Quando la divergenza è sufficientemente antica, le somiglianze osservate tra lingue possono essere dovute prevalentemente a coincidenza stocastica piuttosto che a eredità comune, e nessun algoritmo statistico può distinguere in modo affidabile questi due scenari. Questo pone un limite superiore intrinseco alla profondità temporale che può essere sondata mediante metodi filogenetici basati sul lessico, probabilmente nell'ordine dei diecimila-quindicimila anni nelle condizioni più favorevoli.

Le macrofamiglie ipotizzate: Nostratico, Dené-Caucasico, Austrico

Alcuni linguisti hanno proposto l'esistenza di macrofamiglie che raggrupperebbero diverse famiglie linguistiche ben stabilite in unità genealogiche ancora più ampie, rappresentanti diversificazioni da protolinguaggi parlati nel Paleolitico superiore o addirittura nel Paleolitico medio. Queste proposte rimangono altamente controverse e la maggior parte dei linguisti storici le considera non dimostrate o addirittura metodologicamente infondate.

L'ipotesi nostratica, proposta originariamente dal linguista danese Holger Pedersen nel 1903 e sviluppata estensivamente dalla scuola moscovita di Vladislav Illich-Svitych e Aharon Dolgopolsky, raggrupperebbe l'indoeuropeo, l'uralico, l'altaico, il dravidico, il kartvelo e l'afroasiatico in una superfamiglia parlata ipoteticamente circa quindicimila anni fa. I proponenti hanno compilato liste di centinaia di presunti cognati e hanno tentato ricostruzioni del sistema fonologico del protonostratico.

Tuttavia, la maggior parte dei linguisti storici rigetta queste proposte come metodologicamente difettose. Le principali critiche riguardano l'identificazione eccessivamente permissiva di cognati, che accetta somiglianze fonetiche superficiali senza verificare rigorosamente l'esistenza di corrispondenze sistematiche; la mancanza di ricostruzione dettagliata di morfologia e sintassi, limitandosi principalmente al lessico; e soprattutto il fatto che a profondità temporali di quindicimila anni o più, le somiglianze casuali dovrebbero essere così frequenti da rendere impossibile distinguere eredità genuina da coincidenza.

Gli studi filogenetici computazionali applicati a queste macrofamiglie ipotizzate non hanno prodotto supporto convincente. I valori di verosimiglianza per alberi che raggruppano queste famiglie non sono significativamente superiori a quelli che si otterrebbero per raggruppamenti casuali, e i supporti statistici sono uniformemente bassi. Questo è coerente con l'ipotesi che queste presunte macrofamiglie siano artefatti metodologici piuttosto che entità storiche reali.

Un approccio alternativo e più conservativo consiste nel riconoscere l'esistenza di "lingue isolate" che non possono essere classificate con sicurezza in alcuna famiglia linguistica nota, e nel concentrare gli sforzi di ricerca su famiglie e relazioni che possono essere dimostrate con gli standard rigorosi del metodo comparativo. Il basco, il burushaski, il ainu, il sumerico sono esempi di tali lingue isolate per le quali non è stata stabilita in modo convincente alcuna parentela genealogica con altre lingue.

La ricostruzione fonologica mediante inferenza bayesiana

Un'applicazione particolarmente sofisticata della filogenesi bayesiana consiste nella ricostruzione automatica delle forme ancestrali ai nodi interni dell'albero filogenetico. Tradizionalmente, la ricostruzione comparativa procedeva manualmente, con il linguista che esaminava i riflessi nelle lingue figlie e inferiva la forma più probabile nell'antenato comune applicando il principio della parsimonia e la conoscenza dei pattern universali di cambiamento fonologico.

I metodi bayesiani permettono di formalizzare e automatizzare parzialmente questo processo. Dato un albero filogenetico e un modello di evoluzione fonologica, si può calcolare

la distribuzione a posteriori sulle possibili forme ancestrali a ciascun nodo dell'albero. Il modello di evoluzione fonologica specifica le probabilità relative di diverse transizioni fonetiche: per esempio, una sonorizzazione ($p \rightarrow b$) potrebbe essere più probabile di una desonorizzazione ($b \rightarrow p$), o una palatalizzazione prima di vocale anteriore ($k \rightarrow \check{c} / _i$) più probabile di una palatalizzazione spontanea.

Bouchard-Côté e collaboratori hanno sviluppato modelli sofisticati che incorporano conoscenze fonetiche dettagliate sui meccanismi articolatori e acustici del cambiamento sonoro. Questi modelli trattano ogni segmento fonologico come un fascio di tratti distintivi (sonorità, luogo di articolazione, modo di articolazione) e modellano le transizioni a livello di tratti individuali piuttosto che di fonemi interi. Questo permette di catturare generalizzazioni come il fatto che le occlusive tendono a sonorizzarsi in contesto intersonorico, o che le vocali tendono ad assimilarsi a vocali adiacenti.

L'applicazione di questi metodi alle lingue romanze ha prodotto ricostruzioni del latino volgare che concordano notevolmente bene con le forme attestate nel latino classico e con le ricostruzioni tradizionali dei romanisti. Per esempio, il modello ricostruisce correttamente forme come *porta "porta", *aqua "acqua", *ferrum "ferro" sulla base dei riflessi nelle lingue romanze moderne (italiano "porta", "acqua", "ferro"; francese "porte", "eau", "fer"; spagnolo "puerta", "agua", "hierro").

Tuttavia, l'applicazione a lingue senza attestazioni scritte è più problematica. Le ricostruzioni automatiche tendono a preferire forme fonologicamente semplici e tipologicamente comuni, che potrebbero non riflettere accuratamente la realtà storica se il protolinguaggio aveva effettivamente un sistema fonologico complesso o tipologicamente insolito. Inoltre, i modelli attuali catturano imperfettamente fenomeni come la metatesi, la dissimilazione a distanza, o i cambiamenti morfologicamente condizionati, che sono comuni nella realtà ma difficili da modellare formalmente.

L'applicazione alla ricostruzione semantica e alla paleolinguistica culturale

Oltre alla ricostruzione fonologica, i metodi filogenetici possono essere applicati alla ricostruzione semantica, permettendo di inferire i significati delle parole nel protolinguaggio e quindi di ricostruire aspetti della cultura materiale, dell'organizzazione sociale e dell'ambiente delle popolazioni ancestrali.

Un esempio classico è la ricostruzione del lessico protoindoeuropeo relativo ai mezzi di trasporto e agli animali domestici. L'esistenza di cognati ben stabiliti per "ruota" (**kwekwlo-*), "carro" (**woǵh-*), "asse" (**aks-*), "giogo" (**yugóm*) in quasi tutti i rami della famiglia indoeuropea dimostra che i parlanti del protoindoeuropeo conoscevano la tecnologia del carro a ruote. Poiché questa tecnologia fu inventata approssimativamente nel quarto millennio avanti Cristo nelle steppe euroasiatiche, questo fornisce un limite inferiore per la datazione del protoindoeuropeo e un vincolo sulla sua localizzazione geografica.

Similmente, la ricostruibilità di termini per animali domestici come il cavallo (**h₁ekwos*), la pecora (**h₂ówis*), il bue (**g^wou-*), il maiale (**suH-*) indica un'economia pastorale, mentre la presenza di termini agricoli come "arare" (**h₂erh₃-*), "grano" (**ǵr_hh₂nóm*), "macinare" (**mel_h2-*)

suggerisce anche una componente agricola. Al contrario, l'assenza di cognati sicuri per molti animali e piante tipici del Mediterraneo o del Vicino Oriente (come l'olivo, la vite, l'asino) nelle lingue non meridionali della famiglia suggerisce una patria originaria in regioni più settentrionali.

I metodi bayesiani permettono di formalizzare questo tipo di inferenza. Trattando la presenza o assenza di un termine ricostruibile per un dato concetto come un carattere che evolve lungo l'albero filogenetico, si può stimare la probabilità che quel termine fosse presente in nodi ancestrali specifici. François e collaboratori hanno applicato questo approccio alla ricostruzione del lessico culturale delle lingue austronesiane, stimando le probabilità a posteriori per la presenza di termini relativi a piante domestiche, tecnologie marittime, e organizzazione sociale a diversi livelli della filogenesi austronesiana.

Questo tipo di analisi deve essere interpretato con cautela, perché l'assenza di un cognato ricostruibile non dimostra necessariamente che il concetto fosse assente nella cultura ancestrale: potrebbe semplicemente riflettere sostituzione lessicale indipendente in tutti i rami. Inoltre, i significati delle parole possono cambiare nel tempo, e un cognato apparentemente condiviso potrebbe riflettere non eredità diretta ma sviluppo parallelo da una radice ancestrale con significato diverso. Nonostante queste difficoltà, la paleolinguistica culturale combinata con metodi filogenetici quantitativi offre una finestra unica sulla preistoria culturale umana.

Modelli evolutivi avanzati e analisi di diversificazione

I modelli di speciazione linguistica e i processi di birth-death

La diversificazione delle famiglie linguistiche può essere modellata formalmente attraverso processi stocastici di nascita e morte, analogamente a quanto fatto in biologia evolutiva per la speciazione e l'estinzione. Questi modelli trattano le lingue come entità che possono dividersi (speciazione linguistica), persistere (trasmissione verticale), o estinguersi (morte di una lingua quando la sua ultima comunità di parlanti nativi scompare o passa a un'altra lingua).

Il modello di nascita pura, noto anche come processo di Yule in onore del statistico George Udny Yule che lo formulò nel 1925, rappresenta il caso più semplice. In questo modello, ogni lingua esistente ha una probabilità costante λ per unità di tempo di dividersi in due lingue figlie, e non esiste estinzione. Se partiamo da una singola lingua ancestrale al tempo zero, il numero atteso di lingue al tempo t è $N(t) = \exp(\lambda t)$. Questo processo genera alberi filogenetici ultrametrici (tutti i rami terminali hanno la stessa distanza dalla radice) che crescono esponenzialmente nel tempo.

Il modello nascita-morte è più realistico perché incorpora anche l'estinzione. In questo modello, ogni lingua ha un tasso di speciazione λ e un tasso di estinzione μ . La dinamica dipende crucialmente dal rapporto tra questi due parametri: se $\lambda > \mu$, il numero di lingue cresce nel tempo; se $\lambda = \mu$, il numero rimane stazionario in media; se $\lambda < \mu$, la famiglia si estingue con probabilità uno. La distribuzione del numero di lingue al tempo t e la forma degli alberi filogenetici generati dipendono dalla differenza $\lambda - \mu$ (tasso netto di diversificazione) e dal rapporto μ/λ (frazione di estinzione).

Questi modelli possono essere utilizzati come distribuzioni a priori nella filogenesi bayesiana, generando aspettative probabilistiche sulla forma degli alberi che influenzano poi le inferenze a posteriori. Inoltre, possono essere utilizzati per stimare i tassi di speciazione ed estinzione dai dati filogenetici. Utilizzando metodi di massima verosimiglianza o bayesiani, si può stimare quale combinazione di tassi di nascita e morte spiega meglio la forma osservata dell'albero filogenetico (pattern di ramificazione, distribuzione delle lunghezze dei rami, numero di estinzioni inferite).

Studi che hanno applicato questi modelli alle famiglie linguistiche reali hanno trovato che i tassi di speciazione stimati variano enormemente tra famiglie diverse. Le lingue austronesiane mostrano tassi di speciazione molto alti, coerenti con la loro rapida espansione e diversificazione attraverso il Pacifico negli ultimi cinquemila anni. Le lingue indoeuropee mostrano tassi intermedi, mentre alcune famiglie piccole e geograficamente circoscritte mostrano tassi molto bassi. Questi pattern possono essere correlati con fattori sociodemografici: grandi espansioni geografiche, crescita demografica rapida, e frammentazione in comunità isolate tendono a favorire alti tassi di speciazione linguistica.

La variazione nei tassi di evoluzione e gli orologi molecolari rilassati

Uno dei progressi metodologici più importanti nella filogenesi molecolare degli ultimi vent'anni è stato lo sviluppo di modelli a orologio rilassato che permettono ai tassi di evoluzione di variare tra i diversi rami dell'albero filogenetico. Questo rappresenta un miglioramento cruciale rispetto ai modelli a orologio stretto che assumevano tassi costanti, assunzione che sappiamo essere violata sia in biologia sia, ancora più drasticamente, in linguistica.

Il modello a orologio rilassato lognormale, implementato nel software BEAST, è particolarmente popolare. In questo modello, i tassi di sostituzione lungo i diversi rami dell'albero sono variabili casuali indipendenti estratte da una distribuzione lognormale. La media di questa distribuzione rappresenta il tasso di sostituzione medio dell'intera famiglia, mentre la varianza controlla quanto i tassi possono deviare da questa media. Una varianza bassa corrisponde a un orologio quasi stretto, mentre una varianza alta permette variazioni sostanziali nei tassi tra i rami.

La distribuzione lognormale è stata scelta per ragioni matematiche e biologiche. Matematicamente, essa garantisce che i tassi siano sempre positivi (non può esistere un tasso di sostituzione negativo) ed è conveniente computazionalmente. Biologicamente, molti processi che influenzano i tassi di evoluzione (dimensione della popolazione, tempo di generazione, pressioni selettive) agiscono moltiplicativamente piuttosto che additivamente, e il teorema del limite centrale moltiplicativo implica che l'effetto combinato di molti fattori moltiplicativi indipendenti dovrebbe produrre una distribuzione approssimativamente lognormale.

L'applicazione di questi modelli alle lingue indoeuropee ha rivelato variazioni sostanziali nei tassi di sostituzione lessicale. Le lingue romanze mostrano tassi relativamente alti, particolarmente il francese che ha subito trasformazioni fonologiche e morfologiche drammatiche negli ultimi millecinquecento anni. L'islandese e il lituano mostrano tassi estremamente bassi, avendo conservato caratteristiche arcaiche che le altre lingue

germaniche e baltiche hanno perso. Queste variazioni possono essere correlate con fattori sociolinguistici: presenza o assenza di una forte tradizione letteraria, grado di isolamento geografico, dimensione e mobilità della popolazione parlante.

Un modello alternativo è l'orologio rilassato autocorrelato, nel quale i tassi di sostituzione lungo rami adiacenti dell'albero tendono a essere simili. L'idea è che i fattori che influenzano i tassi di evoluzione (come caratteristiche demografiche o culturali della popolazione) tendono a persistere nel tempo e a essere ereditati dalle lingue figlie, producendo correlazione filogenetica nei tassi. Questo modello può essere più appropriato quando esiste una struttura gerarchica nei fattori che influenzano i tassi, per esempio quando intere famiglie linguistiche condividono caratteristiche socioculturali che influenzano uniformemente i loro tassi di cambiamento.

La calibrazione mediante vincoli fossili linguistici

Analogamente a come i paleontologi utilizzano fossili per calibrare le datazioni filogenetiche in biologia, i linguisti possono utilizzare "fossili linguistici" sotto forma di attestazioni scritte antiche per calibrare gli orologi molecolari linguistici. Tuttavia, l'applicazione pratica di questa procedura richiede considerevole sottigliezza metodologica.

Il concetto di "età minima" e "età massima" per un nodo filogenetico è cruciale. Se una lingua è attestata per iscritto in un determinato periodo, questo fornisce un'età minima sicura per quel ramo dell'albero: sappiamo che la lingua esisteva certamente a quella data. Tuttavia, fornisce solo un limite molto debole sull'età massima: la lingua potrebbe essere molto più antica della sua prima attestazione scritta, perché la scrittura è un'invenzione relativamente recente e la sua adozione per una data lingua dipende da contingenze storiche.

Per i limiti superiori (età massima), spesso dobbiamo ricorrere a ragionamenti più indiretti. Per esempio, se sappiamo da evidenze archeologiche che una certa regione geografica fu colonizzata per la prima volta dagli esseri umani in una data specifica, questo fornisce un limite superiore per l'età delle lingue indigene di quella regione: non possono essere più antiche del primo popolamento umano. Le isole del Pacifico remoto forniscono calibrazioni particolarmente precise di questo tipo, perché la datazione radiocarbonica dei primi insediamenti umani è ben stabilita e fornisce limiti superiori stretti per le lingue polinesiane parlate in quelle isole.

Un altro tipo di calibrazione sfrutta eventi storici documentati di divergenza linguistica. La conquista normanna dell'Inghilterra nel 1066 e il successivo periodo di dominazione franco-normanna forniscono una calibrazione per la divergenza tra il francese continentale e il francese normanno d'Inghilterra (anglo-normanno). La colonizzazione europea delle Americhe a partire dal 1492 fornisce calibrazioni per la divergenza tra le lingue europee e le loro varietà americane. La documentazione storica dell'espansione araba fuori dalla penisola arabica nel settimo secolo fornisce calibrazioni per la diversificazione delle lingue arabe.

Tuttavia, tutti questi eventi forniscono più propriamente vincoli sulla divergenza di dialetti o varietà piuttosto che sulla speciazione di lingue completamente distinte. La transizione da variazione dialettale a lingue separate è graduale e difficile da datare precisamente. Quando esattamente il latino volgare parlato in Gallia cessò di essere "latino" e divenne "francese"?

Non esiste una risposta oggettiva a questa domanda, introducendo ambiguità irriducibile nelle calibrazioni.

Nella pratica bayesiana moderna, queste incertezze vengono incorporate specificando le calibrazioni come distribuzioni probabilistiche piuttosto che come punti fissi. Per esempio, si potrebbe specificare che un nodo di divergenza ha un'età minima di millecinquecento anni con certezza, ma un'età massima descritta da una distribuzione esponenziale con media tremila anni, riflettendo la credenza che età molto maggiori sono possibili ma progressivamente meno probabili. L'analisi bayesiana combina poi queste informazioni a priori con i dati per produrre una distribuzione a posteriori sull'età del nodo che riflette entrambe le fonti di informazione.

Le analisi di diversificazione temporale e i burst di speciazione

Un pattern particolarmente interessante che emerge da molte analisi filogenetiche è l'esistenza di periodi di diversificazione rapida seguiti da periodi di stasi relativa. In biologia evolutiva, questi burst di speciazione sono stati associati a radiazioni adattative, quando un lignaggio colonizza un nuovo ambiente con molte nicchie ecologiche disponibili e si diversifica rapidamente per sfruttarle. Fenomeni analoghi possono verificarsi in linguistica quando una popolazione si espande in un nuovo territorio e si frammenta rapidamente in comunità isolate.

I metodi statistici per rilevare variazioni temporali nei tassi di diversificazione si basano tipicamente sul confronto tra modelli a tasso costante e modelli a tasso variabile. Il test gamma di Pybus e Harvey, originariamente sviluppato per filogenesi molecolari, è stato applicato anche a filogenesi linguistiche. Questo test calcola una statistica gamma che misura se gli eventi di speciazione sono distribuiti uniformemente nel tempo (gamma vicino a zero) o se sono concentrati verso il presente (gamma positivo) o verso il passato vicino alla radice (gamma negativo). Un valore gamma significativamente negativo indica un rallentamento della diversificazione nel tempo, mentre un valore positivo indica un'accelerazione.

Applicato alle lingue indoeuropee, il test gamma ha tipicamente prodotto valori negativi, suggerendo una radiazione rapida iniziale seguita da rallentamento. Questo è coerente con uno scenario nel quale il protoindoeuropeo si frammentò rapidamente in diversi rami principali durante un periodo di espansione geografica, dopo di che i tassi di speciazione diminuirono man mano che le popolazioni si stabilizzarono e i confini linguistici si consolidarono. Pattern simili sono stati osservati in altre famiglie linguistiche associate a espansioni geografiche rapide.

Questi pattern hanno implicazioni importanti per la ricostruibilità della filogenesi profonda. Durante un burst di speciazione rapida, gli intervalli temporali tra eventi di divergenza successivi possono essere molto brevi, dell'ordine di pochi secoli o addirittura decenni. In questo caso, le lingue figlie hanno poco tempo per accumulare innovazioni distintive prima della successiva divergenza, producendo una situazione di politomia quasi irrisolta dove l'ordine esatto delle ramificazioni è difficile o impossibile da determinare con certezza. Questo fenomeno, chiamato "politomia dura" in filogenetica, può spiegare perché certi

aspetti della filogenesi indoeuropea (per esempio, l'ordine esatto di separazione dei rami occidentali) rimangono controversi nonostante decenni di analisi intensiva.

L'inferenza di eventi di estinzione e lingue fantasma

Un aspetto affascinante ma metodologicamente difficile della filogenesi linguistica riguarda l'inferenza di lingue estinte che non hanno lasciato documentazione scritta ma la cui esistenza può essere dedotta dalla struttura dell'albero filogenetico. Queste "lingue fantasma" rappresentano rami dell'albero che si sono estinti completamente, lasciando solo tracce indirette nella topologia filogenetica delle lingue sopravvissute.

Il caso più chiaro riguarda lingue che sappiamo essere esistite da documentazione storica ma che non hanno discendenti diretti. Il latino è l'antenato comune di tutte le lingue romanze, ma molte altre lingue italiche antiche (osco, umbro, falisico, venetico) si estinsero senza lasciare discendenti, anche se sono attestate epigraficamente. Similmente, l'antico prussiano, una lingua baltica documentata fino al diciassettesimo secolo, si estinse completamente senza discendenti moderni.

Più problematico è il caso di lingue completamente non attestate la cui esistenza è inferita indirettamente. La topologia dell'albero filogenetico talvolta implica l'esistenza di nodi ancestrali intermedi tra il protolinguaggio ricostruito e i sottogruppi attestati. Per esempio, l'albero indoeuropeo implica l'esistenza di un proto-germanico che precedette la diversificazione delle lingue germaniche attestate, anche se questa lingua stessa non è direttamente documentata. Similmente, possiamo inferire l'esistenza di un proto-slavo comune, un proto-celtico comune, eccetera.

In alcuni casi, la struttura filogenetica suggerisce l'esistenza di interi rami estinti. Se le lingue di un sottogruppo si collocano filogeneticamente come gruppo fratello di tutte le altre lingue della famiglia messe insieme, questo implica una divergenza molto antica tra quel sottogruppo e il resto della famiglia. Se l'intervallo temporale tra questa divergenza basale e le divergenze successive è molto lungo, ci si aspetterebbe che esistessero altri rami derivanti dal nodo ancestrale che si siano estinti senza lasciare tracce. La loro assenza dai dati attuali riflette semplicemente il fatto che non sono sopravvissuti fino al presente o non sono stati documentati.

I metodi bayesiani permettono di stimare quantitativamente la probabilità e il numero atteso di tali estinzioni. Dato un modello nascita-morte e l'albero osservato delle lingue sopravvissute, si può calcolare la probabilità che esistano rami non campionati (estinti o non documentati) e stimare la loro posizione probabile nell'albero. Queste analisi suggeriscono tipicamente che una frazione sostanziale della diversità linguistica originale si è estinta senza lasciare tracce, particolarmente per famiglie con grande profondità temporale.

La correzione per bias di campionamento e lingue non rappresentate

Un problema metodologico spesso trascurato nella filogenesi linguistica riguarda il fatto che i dataset disponibili non rappresentano un campionamento casuale di tutte le lingue della famiglia, ma sono fortemente biased verso lingue ben documentate, lingue con molti

parlanti, e lingue di aree geografiche ben studiate. Questo bias di campionamento può distorcere le inferenze filogenetiche in modi sottili ma importanti.

Per le famiglie linguistiche moderne con documentazione relativamente completa, molte varietà dialettali intermedie tra le lingue standard riconosciute sono escluse dai dataset per ragioni pratiche. Per esempio, gli studi filogenetici sulle lingue romanze tipicamente includono le lingue letterarie standard (italiano, francese, spagnolo, portoghese, rumeno) ma escludono le centinaia di varietà dialettali locali che esistono in ogni regione del territorio romanzo. Questo crea un'impressione artificiosa di divergenza netta tra entità discrete, quando la realtà storica era probabilmente un continuum dialettale con transizioni graduali.

Inoltre, molte lingue si sono estinte storicamente a causa della colonizzazione, della sostituzione linguistica, o semplicemente della piccola dimensione delle loro comunità parlanti, e non appaiono nei dataset moderni. Le famiglie linguistiche native delle Americhe hanno subito estinzioni massive a seguito della colonizzazione europea, con stime che suggeriscono che potrebbero essere scomparse più del cinquanta per cento delle lingue amerindie originarie. Similmente, molte lingue dell'Australia aborigena si sono estinte o sono criticamente in pericolo.

Queste estinzioni non casuali possono introdurre bias sistematici. Se le lingue più divergenti o quelle in regioni geograficamente remote hanno maggiore probabilità di estinguersi senza essere documentate, gli alberi filogenetici ricostruiti dai dati sopravvissuti sottostimeranno la diversità originale e potrebbero dare un'immagine distorta dei pattern di diversificazione. Metodi statistici per correggere il bias di campionamento esistono, ma richiedono assunzioni su quale frazione delle lingue originali è stata campionata e su se il campionamento è casuale o sistematicamente biased, assunzioni che spesso sono difficili da giustificare.

Una strategia pratica consiste nel testare la robustezza delle conclusioni rispetto all'inclusione o esclusione di sottoinsiemi dei dati. Se le conclusioni principali (topologia dell'albero, datazioni dei nodi maggiori, tassi di diversificazione) rimangono stabili quando si analizzano solo le lingue meglio documentate versus quando si include la massima copertura possibile includendo lingue più scarsamente documentate, questo aumenta la confidenza che il bias di campionamento non sta distorcendo gravemente le inferenze.

Confronti interdisciplinari e convergenza di evidenze

La correlazione tra filogenesi linguistica e genetica delle popolazioni

Una delle applicazioni più affascinanti della filogenesi linguistica computazionale riguarda il confronto tra alberi filogenetici linguistici e alberi filogenetici genetici delle popolazioni che parlano quelle lingue. L'idea sottostante è semplice: se le lingue si trasmettono prevalentemente verticalmente da genitori a figli insieme al materiale genetico, ci si aspetta che i pattern di diversità linguistica e genetica dovrebbero essere correlati, riflettendo entrambi la storia demografica delle popolazioni.

Il genetista delle popolazioni Luigi Luca Cavalli-Sforza fu pioniere di questo approccio negli anni Settanta e Ottanta, correlando distanze genetiche basate su frequenze alleliche con distanze linguistiche tra le popolazioni. I suoi studi rivelarono correlazioni significative, anche

se imperfette, tra diversità genetica e linguistica, particolarmente a livello di grandi famiglie linguistiche e gruppi continentali. Per esempio, la principale divisione genetica nell'umanità, quella tra popolazioni africane e non-africane, corrisponde approssimativamente a una divisione linguistica maggiore, anche se con molte eccezioni dovute a migrazioni e sostituzioni linguistiche.

L'avvento delle tecnologie di sequenziamento genomico ad alta efficienza ha permesso analisi molto più dettagliate. Gli studi moderni utilizzano tipicamente marcatori genetici ad alta risoluzione come i polimorfismi di singolo nucleotide (SNP) distribuiti attraverso l'intero genoma, o sequenze complete del DNA mitocondriale e del cromosoma Y, che vengono ereditati rispettivamente per via materna e paterna e quindi tracciano linee di discendenza specifiche.

Per le popolazioni austronesiane, studi genetici hanno confermato in modo impressionante le ricostruzioni linguistiche. Le popolazioni di lingua austronesiana dell'Asia sudorientale insulare e del Pacifico mostrano una componente genetica distintiva che le differenzia dalle popolazioni indigene della Nuova Guinea e dell'Australia, riflettendo un'espansione demografica relativamente recente da Taiwan attraverso le Filippine e l'Indonesia. La diversità genetica diminuisce progressivamente da ovest verso est attraverso il Pacifico, con le popolazioni polinesiane più orientali che mostrano la diversità più bassa, coerente con ripetuti effetti di fondazione durante la colonizzazione progressiva delle isole remote. Questo pattern corrisponde quasi perfettamente al pattern di diversità linguistica rivelato dalla filogenesi computazionale.

Per le popolazioni indoeuropee, la situazione è più complessa ma ugualmente illuminante. Le analisi genomiche recenti hanno rivelato che le popolazioni europee moderne derivano principalmente da tre componenti ancestrali: cacciatori-raccoglitori mesolitici indigeni, agricoltori neolitici del Vicino Oriente, e pastori delle steppe dell'età del bronzo associati alla cultura Yamnaya. Quest'ultima componente, che raggiunge il massimo nelle popolazioni dell'Europa settentrionale e orientale e diminuisce verso sud e ovest, è stata interpretata come la firma genetica dell'espansione dei popoli indoeuropei dalle steppe ponto-caspiche circa cinquemila anni fa.

Crucialmente, le datazioni genetiche per questo evento di mescolanza (circa 3000-2500 avanti Cristo) concordano bene con le datazioni archeologiche per l'apparizione di nuove culture materiali in Europa centrale associate ai popoli delle steppe, e con le datazioni linguistiche dell'ipotesi delle steppe per l'espansione indoeuropea, ma non con le datazioni molto più antiche (8000-9000 avanti Cristo) proposte dagli studi bayesiani che supportano l'ipotesi anatolica. Questa discrepanza ha portato molti ricercatori a concludere che le datazioni linguistiche bayesiane molto antiche riflettono probabilmente problemi metodologici piuttosto che la realtà storica.

Le discordanze tra filogenesi linguistica e genetica: sostituzione linguistica e contatto

Nonostante la correlazione generale tra diversità linguistica e genetica, esistono casi ben documentati di discordanza drammatica che rivelano processi storici complessi di sostituzione linguistica senza sostituzione genetica significativa, o viceversa.

L'Ungheria fornisce un esempio classico. Gli ungheresi parlano una lingua uralica (il magiario) che è genealogicamente imparentata con il finlandese e le lingue samoiede della Siberia, lingue geograficamente remote e tipologicamente molto diverse dalle lingue indoeuropee circostanti. Tuttavia, geneticamente gli ungheresi moderni sono indistinguibili dalle altre popolazioni dell'Europa centrale e orientale. Studi sul DNA antico hanno rivelato che i magiari che invasero il bacino carpatico nel nono secolo dell'era comune erano una popolazione relativamente piccola di pastori nomadi delle steppe che impose la sua lingua a una popolazione agricola preesistente molto più numerosa, producendo sostituzione linguistica senza sostituzione genetica corrispondente. Le analisi genetiche suggeriscono che meno del dieci per cento dell'ascendenza genetica degli ungheresi moderni deriva dai magiari invasori originari.

Un caso ancora più estremo è rappresentato dalla Turchia. I turchi parlano una lingua turca appartenente alla famiglia altaica, con affinità genealogiche verso le lingue dell'Asia centrale. Tuttavia, geneticamente le popolazioni dell'Anatolia moderna sono essenzialmente identiche alle popolazioni pre-turche della regione, con contributi genetici minimi dalle steppe dell'Asia centrale. Le invasioni turche selgiuchide e poi ottomane portarono all'adozione della lingua turca da parte delle popolazioni anatoliche preesistenti (che probabilmente parlavano greco, armeno, e altre lingue locali) senza sostituzione demografica significativa.

Questi casi dimostrano che la lingua e i geni possono dissociarsi completamente quando piccole élite politicamente dominanti impongono la loro lingua a popolazioni conquistate molto più numerose. Altri fattori che possono produrre dissociazioni includono l'adozione volontaria di lingue prestigiose per mobilità sociale (come l'adozione del latino da parte delle élite galliche romanizzate), o le politiche linguistiche statali che promuovono una lingua standard a scapito di varietà locali.

Esistono anche casi opposti di sostituzione genetica senza sostituzione linguistica corrispondente. Le popolazioni di lingua bantu dell'Africa meridionale, per esempio, parlano lingue chiaramente imparentate con le lingue bantu dell'Africa occidentale e centrale, ma mostrano componenti genetiche sostanziali derivate dalle popolazioni Khoisan indigene dell'Africa meridionale, riflettendo mescolanza significativa durante l'espansione bantu. In questo caso, l'espansione demografica dei popoli bantu fu accompagnata da assorbimento genetico di popolazioni indigene che adottarono le lingue e le culture bantu.

Questi pattern complessi sottolineano che, mentre la trasmissione verticale congiunta di lingue e geni rappresenta la situazione di default, i processi storici reali sono spesso molto più complessi, coinvolgendo mescolanze, sostituzioni, e adozioni che possono dissociare parzialmente o completamente le storie linguistiche e genetiche.

L'integrazione con dati archeologici: cultura materiale e espansioni linguistiche

L'archeologia fornisce una terza linea di evidenza indipendente che può essere integrata con le ricostruzioni linguistiche e genetiche per costruire narrazioni storiche più complete. Le culture materiali identificate archeologicamente (ceramiche, utensili, armi, abitazioni, pratiche funerarie) possono essere correlate con espansioni linguistiche quando esistono

discontinuità archeologiche nette che coincidono temporalmente e geograficamente con le espansioni inferite linguisticamente.

L'espansione delle lingue bantu attraverso l'Africa subsahariana è forse l'esempio meglio documentato di integrazione tra evidenze linguistiche, genetiche e archeologiche. La filogenesi linguistica indica un'origine delle lingue bantu nel Camerun occidentale circa cinquemila anni fa, seguita da un'espansione verso est e sud che culminò con la colonizzazione dell'Africa meridionale circa duemila anni fa. Questa ricostruzione corrisponde notevolmente bene con l'evidenza archeologica per la diffusione dell'agricoltura basata su igname e banane, della metallurgia del ferro, e di tradizioni ceramiche distintive attraverso la stessa regione e nello stesso arco temporale.

Gli studi genetici sul DNA delle popolazioni bantu moderne e sul DNA antico estratto da scheletri archeologici confermano un'espansione demografica da ovest verso est e sud, con le popolazioni bantu dell'Africa meridionale che mostrano ridotta diversità genetica coerente con ripetuti effetti di fondazione. La convergenza di queste tre linee di evidenza indipendenti fornisce un supporto molto robusto alla ricostruzione storica complessiva dell'espansione bantu.

Per le lingue indoeuropee, l'integrazione con l'archeologia è più controversa ma ugualmente cruciale. L'ipotesi delle steppe associa l'espansione indoeuropea con la cultura Yamnaya delle steppe ponto-caspiche circa 3000 avanti Cristo, caratterizzata da pastorizia mobile, uso del cavallo domestico, carri a ruote, sepolture in kurgan (tumuli), e una serie di innovazioni tecnologiche e sociali. L'evidenza archeologica mostra l'apparizione improvvisa di elementi culturali derivati dalle steppe in Europa centrale e settentrionale circa 2900-2500 avanti Cristo (cultura della Ceramica Cordata), coincidente con l'arrivo della componente genetica delle steppe rivelata dagli studi sul DNA antico.

Al contrario, l'ipotesi anatolica associa l'espansione indoeuropea con la diffusione dell'agricoltura neolitica dall'Anatolia verso l'Europa circa 7000-6000 avanti Cristo. L'evidenza archeologica per questa espansione neolitica è solida, ma gli studi genetici indicano che la componente genetica neolitica dall'Anatolia fu parzialmente sostituita dalla successiva componente delle steppe nell'età del bronzo. Questo ha portato molti ricercatori a proporre scenari complessi con multiple ondate di espansione: forse i primi agricoltori neolitici parlavano lingue indoeuropee "proto-anatoliche" che furono poi sostituite in gran parte dell'Europa dalle lingue indoeuropee "classiche" portate dai popoli delle steppe, con il ramo anatolico (ittita) che rappresenta un relitto della prima espansione.

Queste interpretazioni rimangono controverse, illustrando sia il potere sia i limiti dell'integrazione interdisciplinare. Quando evidenze da discipline diverse convergono, come nel caso bantu, la confidenza nelle ricostruzioni storiche è notevolmente rafforzata. Quando le evidenze sono discordanti o ambigue, come nel caso indoeuropeo, l'interpretazione richiede cautela e il riconoscimento che scenari storici multipli possono essere compatibili con i dati disponibili.

La paleoclimatologia e i vincoli ambientali sulle ricostruzioni linguistiche

Una linea di evidenza più indiretta ma potenzialmente informativa viene dalla paleoclimatologia, lo studio dei climi passati attraverso proxy geologici e biologici. I

cambiamenti climatici hanno influenzato profondamente le popolazioni umane, causando migrazioni, espansioni, collassi demografici, e trasformazioni culturali che dovrebbero aver lasciato tracce nelle storie linguistiche.

Il periodo caldo dell'Olocene medio (approssimativamente 8000-5000 avanti Cristo) vide condizioni climatiche generalmente favorevoli nell'Eurasia temperata, con temperature più calde dell'attuale e precipitazioni più abbondanti. Questo periodo coincise con la rivoluzione neolitica e l'espansione dell'agricoltura fuori dal Vicino Oriente, eventi che probabilmente facilitarono anche espansioni linguistiche. La successiva fase più fredda e arida dell'Olocene tardo (dopo 3000 avanti Cristo) potrebbe aver spinto popolazioni pastorali delle steppe verso regioni più favorevoli, contribuendo alle migrazioni indoeuropee.

Su scale temporali più brevi, eventi climatici drammatici come la crisi del 4200 avanti Cristo (un periodo di aridità estrema nel Vicino Oriente e nel Mediterraneo) o il crollo dell'età del bronzo circa 1200 avanti Cristo (associato a siccità prolungate e carestie) potrebbero aver causato frammentazioni demografiche e movimenti di popolazioni che influenzarono le storie linguistiche. Tuttavia, stabilire collegamenti causali diretti tra eventi climatici e cambiamenti linguistici è metodologicamente difficile, richiedendo correlazioni temporali precise che spesso non sono disponibili data l'incertezza nelle datazioni linguistiche.

Un'applicazione più promettente della paleoclimatologia riguarda la ricostruzione del lessico proto-ambientale. Se possiamo ricostruire termini per piante, animali, e fenomeni climatici nel protolinguaggio, questo fornisce vincoli sulla gamma di ambienti che i parlanti devono aver abitato. Per esempio, l'esistenza di termini ricostruibili per alberi temperati come la betulla, il faggio, e la quercia nel protoindoeuropeo, ma l'assenza di termini per piante tropicali, suggerisce una localizzazione in zone temperate. La ricostruibilità di termini per la neve e il ghiaccio suggerisce che i parlanti sperimentavano inverni freddi. Confrontando queste inferenze con le ricostruzioni paleoclimatiche delle condizioni ambientali in diverse regioni e periodi, si possono porre vincoli geografici e temporali sulle patrie originarie delle famiglie linguistiche.

Le simulazioni computazionali dell'evoluzione linguistica

Un approccio metodologico relativamente nuovo consiste nell'utilizzare simulazioni al computer per modellare esplicitamente i processi di evoluzione linguistica e testare se i pattern osservati nei dati reali sono compatibili con determinate ipotesi sui meccanismi sottostanti. Questi modelli "agent-based" (basati su agenti) simulano popolazioni di parlanti individuali che interagiscono socialmente, trasmettono la lingua ai loro figli, migrano geograficamente, e innovano linguisticamente secondo regole probabilistiche specificate.

Un vantaggio cruciale delle simulazioni è che permettono di esplorare controfattuali che sarebbero impossibili da testare empiricamente. Per esempio, si può simulare cosa succederebbe alla diversità linguistica se si variassero sistematicamente parametri come il tasso di innovazione lessicale, la mobilità geografica delle popolazioni, l'intensità del contatto tra comunità adiacenti, o la pressione verso conformità linguistica all'interno delle comunità. Confrontando i pattern emergenti dalle simulazioni con i pattern osservati nei dati reali, si può inferire quali combinazioni di parametri sono più plausibili.

James Steele e colleghi hanno utilizzato simulazioni di questo tipo per esplorare l'espansione delle lingue bantu. I loro modelli incorporano crescita demografica, dispersione geografica limitata da barriere ambientali, e processi di speciazione linguistica quando popolazioni si separano geograficamente. I risultati mostrano che pattern di diversità linguistica simili a quelli osservati nelle lingue bantu reali emergono naturalmente quando si assumono tassi di crescita demografica e mobilità geografica coerenti con popolazioni agricole in espansione.

Altri ricercatori hanno utilizzato simulazioni per esplorare come il contatto linguistico e il prestito influenzano la ricostruibilità della filogenesi. Questi studi mostrano che anche livelli moderati di prestito lessicale possono distorcere significativamente le topologie filogenetiche ricostruite con metodi standard, producendo raggruppamenti spuri che riflettono contatto geografico piuttosto che eredità comune. Tuttavia, metodi che modellano esplicitamente il contatto (come gli approcci a rete discussi precedentemente) sono più robusti a questo tipo di interferenza.

Le simulazioni possono anche testare l'effetto del campionamento incompleto. Simulando l'evoluzione completa di una famiglia linguistica e poi "campionando" solo una frazione delle lingue risultanti (analogamente a come i dataset reali campionano in modo incompleto la diversità linguistica totale), si può valutare quanto le topologie e le datazioni ricostruite dai dati campionati deviano dalla verità simulata conosciuta. Questi studi suggeriscono generalmente che il campionamento del venti-trenta per cento delle lingue è sufficiente per ricostruire accuratamente la topologia maggiore dell'albero, ma che le datazioni assolute possono essere significativamente distorte se il campionamento è geograficamente o temporalmente biased.

I database lessicali computazionali e il progetto Lexibank

Lo sviluppo di database lessicali standardizzati, computazionalmente accessibili, e verificati per qualità rappresenta un progresso infrastrutturale cruciale per la filogenesi linguistica comparativa. Il progetto Lexibank, coordinato dal Max Planck Institute for Evolutionary Anthropology di Lipsia, aggrega dataset lessicali per centinaia di famiglie linguistiche in un formato standardizzato che facilita l'analisi computazionale su larga scala.

Lexibank non è semplicemente una collezione di liste lessicali, ma un sistema curato che include metadati dettagliati sulle lingue (affiliazione genealogica, localizzazione geografica, numero di parlanti, stato di documentazione), forme fonetiche trascritte in alfabeto fonetico internazionale standardizzato, identificazioni di cognati verificate da esperti, e collegamenti a fonti bibliografiche. Questa strutturazione rigorosa permette analisi comparative che sarebbero impraticabili con dati meno organizzati.

Un aspetto particolarmente importante è la codifica esplicita dell'incertezza. Per ogni giudizio di cognazione (due forme sono cognate o no?), Lexibank permette di specificare livelli di confidenza, distinguendo cognati certi da cognati probabili da cognati possibili. Questo permette analisi di sensibilità dove si testano come le conclusioni cambiano quando si includono solo cognati certi versus quando si includono anche cognati meno sicuri.

Il database include anche informazioni sulla fonte dei dati e sulla metodologia utilizzata per la loro raccolta. Questo è cruciale perché la qualità dei dati varia enormemente: liste

compile da linguisti specialisti sul campo usando metodologie rigorose sono molto più affidabili di liste estratte da dizionari missionari ottocenteschi o compilate da non-specialisti. Esplicitando queste informazioni di provenienza, Lexibank permette agli utenti di valutare criticamente la qualità dei dati e di pesare appropriatamente le diverse fonti nelle analisi.

L'esistenza di database come Lexibank facilita anche la replicabilità, un aspetto cruciale del metodo scientifico che è stato spesso trascurato nella linguistica storica tradizionale. Quando i dataset e gli script di analisi sono resi pubblicamente disponibili in formati standardizzati, altri ricercatori possono ripetere le analisi, testare ipotesi alternative, o estendere gli studi a nuove lingue, accelerando il progresso cumulativo della conoscenza.

Questioni teoriche avanzate e prospettive epistemologiche

Il problema dell'identificabilità dei modelli e i limiti della risoluzione filogenetica

Un problema fondamentale nella filogenesi computazionale riguarda l'identificabilità: dati i dati osservati, esiste una combinazione unica di parametri del modello (topologia dell'albero, lunghezze dei rami, tassi di sostituzione) che spiega ottimamente quei dati, oppure esistono multiple soluzioni equivalenti o quasi-equivalenti che non possono essere distinte statisticamente? Questo problema ha implicazioni profonde per l'interpretabilità delle ricostruzioni filogenetiche.

Dal punto di vista della teoria dell'informazione, la quantità di informazione filogenetica contenuta in un dataset è limitata. Per caratteri binari come la presenza/assenza di cognati, il contenuto informativo massimo per carattere è un bit. Un dataset con mille caratteri può quindi contenere al massimo mille bit di informazione filogenetica, anche se nella pratica il contenuto effettivo è molto inferiore a causa di correlazioni tra caratteri, rumore, e omoplasia (evoluzione convergente).

D'altra parte, la complessità dello spazio degli alberi filogenetici cresce estremamente rapidamente con il numero di lingue. Per n lingue, esistono $(2n-3)!! = (2n-3) \times (2n-5) \times \dots \times 3 \times 1$ alberi radicati binari possibili. Per dieci lingue, questo numero è già oltre due milioni; per venti lingue, supera otto per dieci alla potenza venti. La maggior parte di questi alberi può essere esclusa rapidamente perché ha verosimiglianza bassissima data i dati, ma tipicamente rimangono molti alberi con verosimiglianze simili che non possono essere distinti in modo affidabile.

Questo problema è particolarmente acuto per eventi di diversificazione rapida (radiazioni adattative) dove molteplici speciazioni avvengono in intervalli temporali brevi. In questo scenario, i rami interni dell'albero che connettono questi eventi sono molto corti, e le lingue risultanti hanno poco tempo per accumulare innovazioni distintive prima della successiva biforcazione. Di conseguenza, l'ordine esatto delle ramificazioni diventa statisticamente irrisolvibile: i dati sono compatibili con multiple topologie alternative che differiscono solo nell'ordine di eventi di speciazione ravvicinati.

I metodi bayesiani hanno il vantaggio di quantificare esplicitamente questa incertezza attraverso le probabilità a posteriori. Se la probabilità a posteriori per una particolare risoluzione di un nodo è cinquanta per cento mentre una risoluzione alternativa ha quarantacinque per cento, questo segnala che i dati non permettono di distinguere con confidenza tra le due possibilità. Interpretazioni che trattano una di queste risoluzioni come "vera" mentre l'altra è "falsa" sarebbero epistemologicamente ingiustificate.

Questa incertezza irriducibile non riflette necessariamente inadeguatezza dei metodi o dei dati, ma può riflettere l'effettiva struttura della realtà storica. Se il protoindoeuropeo si frammentò effettivamente in diversi rami principali nell'arco di pochi secoli durante un'espansione geografica rapida, l'ordine preciso di queste frammentazioni potrebbe non aver lasciato tracce filogenetiche distinguibili, e l'ambiguità osservata nelle ricostruzioni riflette fedelmente questa realtà storica.

La tensione tra modelli evolutivi neutri e adattativi

La filogenesi molecolare in biologia si basa largamente sulla teoria neutrale dell'evoluzione molecolare, che assume che la maggior parte delle mutazioni a livello del DNA sono selettivamente neutre e si fissano nelle popolazioni per deriva genetica casuale. Questo permette di utilizzare modelli stocastici relativamente semplici di mutazione e deriva per modellare l'evoluzione delle sequenze.

Nella linguistica, tuttavia, la situazione è molto più complessa. Mentre alcuni aspetti del cambiamento linguistico possono essere approssimativamente neutri (per esempio, variazioni fonetiche minori che non influenzano la comprensibilità), molti cambiamenti hanno conseguenze funzionali e comunicative che potrebbero essere soggette a qualcosa di analogo alla selezione naturale. Innovazioni che migliorano l'espressività, riducono l'ambiguità, facilitano l'apprendimento, o aumentano il prestigio sociale potrebbero diffondersi preferenzialmente, mentre innovazioni disfunzionali potrebbero essere selezionate contro.

Per esempio, la perdita di distinzioni fonologiche che creano troppi omofoni (parole foneticamente identiche con significati diversi) potrebbe essere selezionata contro perché riduce l'efficacia comunicativa. Al contrario, semplificazioni morfologiche che riducono la ridondanza senza compromettere la comunicazione potrebbero essere favorite perché facilitano l'apprendimento. Questi processi "adattativi" violano l'assunzione di neutralità che sottende molti modelli filogenetici standard.

La questione cruciale è: quanto importante è questa violazione per l'accuratezza delle ricostruzioni filogenetiche? Se la maggior parte dei cambiamenti lessicali nel vocabolario di base sono approssimativamente neutri, i modelli neutri potrebbero fornire approssimazioni ragionevoli nonostante violazioni occasionali. Tuttavia, se esistono bias sistematici verso certi tipi di cambiamenti, questi potrebbero distorcere le ricostruzioni filogenetiche in modi prevedibili.

Alcuni ricercatori hanno tentato di sviluppare modelli esplicitamente adattativi che incorporano pressioni selettive. Per esempio, si potrebbe modellare la probabilità che un termine lessicale venga sostituito come funzione della sua frequenza d'uso, basandosi sull'osservazione empirica che parole molto frequenti tendono ad essere più conservate.

Oppure si potrebbero incorporare vincoli tipologici che rendono certi cambiamenti strutturali più probabili di altri a seconda delle caratteristiche esistenti della lingua.

Tuttavia, questi modelli adattativi sono molto più complessi e richiedono la stima di parametri aggiuntivi, riducendo il potere statistico delle analisi. Inoltre, le "leggi" che governano il cambiamento linguistico sono molto meno ben comprese rispetto alle leggi della genetica di popolazione, rendendo difficile specificare modelli adattativi ben motivati. Per questi motivi, la maggior parte delle applicazioni attuali continua a utilizzare modelli neutrali o quasi-neutrali, riconoscendo che rappresentano semplificazioni ma sperando che siano approssimazioni di primo ordine ragionevoli.

La questione ontologica: cosa sono le lingue?

Un problema filosofico sottostante a tutta la filogenesi linguistica riguarda la definizione stessa di "lingua" come entità discreta soggetta a relazioni filogenetiche. Nella realtà sociolinguistica, quello che esiste sono parlanti individuali con repertori linguistici che variano continuamente nello spazio geografico e sociale. I confini tra "lingue" diverse sono spesso vaghi, arbitrari, e determinati più da fattori politici che linguistici.

Il caso paradigmatico è il continuum dialettale, dove varietà adiacenti sono mutualmente intelligibili ma varietà agli estremi opposti del continuum non lo sono. L'Europa occidentale prima della standardizzazione linguistica moderna costituiva essenzialmente un continuum romanzo che si estendeva dal Portogallo alla Romania, con transizioni graduali piuttosto che confini netti. Parlanti di villaggi adiacenti potevano comunicare senza difficoltà, ma parlanti di regioni distanti non potevano. Dove tracciare i confini tra "italiano", "francese", "spagnolo" in questa situazione?

La risposta storica è stata largamente politica: i confini linguistici furono in gran parte creati dai processi di standardizzazione e nazionalizzazione dei secoli diciottesimo e diciannovesimo, quando stati emergenti promossero dialetti specifici (tipicamente quelli delle capitali politiche) come lingue nazionali standard, marginalizzando gli altri dialetti. Ma questa categorizzazione riflette contingenze politiche piuttosto che discontinuità linguistiche naturali.

Questo problema ontologico complica la filogenesi linguistica computazionale perché i metodi assumono l'esistenza di entità discrete ("lingue") che possono essere posizionate come foglie terminali di un albero filogenetico. Se nella realtà esistono continua dialettali fluidi piuttosto che entità discrete, l'intera impresa di costruire alberi filogenetici potrebbe essere mal posta.

Una risposta pragmatica è riconoscere che le "lingue" nei dataset filogenetici sono costrutti analitici piuttosto che entità naturali, ma che possono comunque essere utili per tracciare pattern storici. Finché le varietà campionate rappresentano approssimativamente punti discreti e ben separati nello spazio delle variazioni linguistiche, le relazioni filogenetiche tra di esse possono essere significative anche se non catturano la complessità totale della variazione reale.

Un'altra risposta è sviluppare metodi che rappresentano esplicitamente la continuità dialettale piuttosto che assumere entità discrete. I modelli a rete discussi precedentemente rappresentano un passo in questa direzione, permettendo rappresentazioni più flessibili che

catturano mescolanze e transizioni graduali. Metodi ancora più sofisticati potrebbero modellare le lingue come distribuzioni in spazi continui di variazione piuttosto che come punti discreti, approssimando meglio la realtà sociolinguistica.

Il problema della circolarità nelle datazioni e nelle calibrazioni

Un problema metodologico sottile ma importante riguarda la potenziale circolarità tra le inferenze filogenetiche linguistiche e le interpretazioni archeologiche o genetiche usate per calibrarle o validarle. Se usiamo una ipotesi archeologica (per esempio, che una certa espansione culturale riflette una espansione linguistica) per calibrare un'analisi filogenetica, e poi usiamo i risultati di quell'analisi per supportare la stessa ipotesi archeologica, stiamo ragionando circolarmente.

Per esempio, se assumiamo che la cultura Yamnaya rappresenta l'espansione indoeuropea e usiamo la datazione archeologica di questa cultura (circa 3000 avanti Cristo) per calibrare il nodo protoindoeuropeo nell'analisi filogenetica, e poi concludiamo che l'analisi supporta una datazione del protoindoeuropeo intorno al 3000 avanti Cristo coerente con l'ipotesi Yamnaya, non abbiamo realmente testato l'ipotesi in modo indipendente ma l'abbiamo incorporata nelle assunzioni a priori dell'analisi.

Evitare questa circolarità richiede cautela metodologica. Idealmente, le calibrazioni dovrebbero basarsi su evidenze storiche indipendenti e incontroverse (come le prime attestazioni scritte), non su ipotesi archeologiche controverse che sono esse stesse oggetto di verifica. Quando si utilizzano calibrazioni archeologiche, dovrebbe essere esplicitato chiaramente che rappresentano assunzioni ausiliarie, e dovrebbe essere testato quanto i risultati sono sensibili a calibrazioni alternative.

Un approccio più rigoroso consiste nell'utilizzare metodi di inferenza bayesiana gerarchica che trattano anche le calibrazioni come variabili incerte piuttosto che come input fissi. In questo framework, si specificano distribuzioni a priori non solo sui parametri del modello filogenetico ma anche sui parametri che controllano le calibrazioni, e si lascia che i dati informino congiuntamente tutti questi parametri. Questo può ridurre (anche se non eliminare completamente) i problemi di circolarità permettendo ai dati di "parlare" più liberamente.

La questione della validità ecologica dei modelli semplificati

Tutti i modelli filogenetici attuali, per quanto matematicamente sofisticati, rappresentano semplificazioni drastiche della complessità reale del cambiamento linguistico. Tipicamente assumono che le lingue possano essere rappresentate come collezioni di caratteri discreti (cognati presenti/assenti, caratteristiche strutturali presenti/assenti) che evolvono indipendentemente secondo processi stocastici semplici. Ma le lingue reali sono sistemi strutturati complessi dove i componenti interagiscono in modi non-lineari e dove i cambiamenti a un livello (per esempio, fonologia) possono cascadere a altri livelli (morfologia, sintassi, lessico).

Per esempio, un cambiamento fonologico che introduce una nuova distinzione tonale può innescare ristrutturazioni morfologiche che utilizzano i toni per distinguere forme grammaticali, che a loro volta possono influenzare l'ordine delle parole e la struttura della

frase. Questi effetti a cascata violano l'assunzione di indipendenza dei caratteri e creano correlazioni complesse che i modelli standard non catturano.

Similmente, i modelli attuali tipicamente ignorano la variazione intralinguistica (dialettale, sociale, stilistica) che è onnipresente nelle lingue reali. Trattano ogni lingua come un'entità omogenea con un insieme fisso di caratteristiche, quando in realtà ogni lingua esiste come un insieme di varietà in variazione. Questa semplificazione può introdurre artefatti quando lingue con alta variazione interna vengono campionate da parlanti o regioni specifiche che non rappresentano l'intera gamma di variazione.

La domanda cruciale è: nonostante queste semplificazioni drastiche, i modelli catturano abbastanza della dinamica reale per produrre inferenze approssimativamente corrette, oppure le violazioni delle assunzioni sono così severe da rendere le inferenze inaffidabili? Non esiste una risposta definitiva a questa domanda, e probabilmente la risposta varia da caso a caso.

La strategia pragmatica è continuare a utilizzare modelli semplificati riconoscendone i limiti, mentre simultaneamente si lavora allo sviluppo di modelli più realistici che catturano più della complessità reale. La validazione su casi noti (come le lingue romanze) suggerisce che almeno in alcuni casi i modelli semplificati producono risultati ragionevolmente accurati, fornendo fiducia che non sono completamente inadeguati. Tuttavia, prudenza è necessaria quando si estrapolano risultati a casi meno ben compresi.

La dimensione sociale e ideologica della ricerca filogenetica

La ricerca sulla preistoria linguistica non avviene in un vuoto sociale e politico ma è intrecciata con questioni di identità nazionale, indigenità, e diritti territoriali che possono avere conseguenze pratiche significative. Le conclusioni sulla parentela linguistica, le patrie originarie, e le rotte migratorie possono essere utilizzate (o abusate) per supportare rivendicazioni politiche o per costruire narrazioni nazionali.

Per esempio, le controversie sulla relazione tra le lingue sinotibetane e le loro implicazioni per la storia antica della Cina hanno dimensioni politiche perché toccano questioni sensibili riguardo alle relazioni tra l'etnia Han maggioritaria e le minoranze etniche nelle regioni periferiche della Cina. Similmente, le ricerche sull'origine e la diffusione delle lingue indoeuropee toccano questioni di identità nazionale europea e possono essere strumentalizzate da movimenti nazionalisti.

Gli studiosi hanno la responsabilità di essere consapevoli di queste dimensioni sociali e di comunicare i loro risultati in modi che minimizzino il rischio di appropriazione indebita. Questo include essere espliciti sui limiti e sulle incertezze delle ricostruzioni, resistere a interpretazioni deterministiche o essenzialiste delle identità linguistiche, e contestualizzare i risultati scientifici in quadri interpretativi più ampi che riconoscono la complessità della storia umana.

È anche importante riconoscere che le categorie e le tassonomie utilizzate nella ricerca linguistica non sono neutre ma riflettono storicamente prospettive coloniali e eurocentriche. Molte delle classificazioni linguistiche ancora in uso furono sviluppate nel contesto dell'imperialismo europeo e riflettono assunzioni problematiche sulla superiorità di certe

lingue o culture. La decolonizzazione della linguistica richiede un riesame critico di queste categorie e un maggiore coinvolgimento delle comunità linguistiche stesse nella ricerca che le riguarda.

Verso una sintesi: filogenetica come componente di un approccio pluralistico

In conclusione, la filogenesi linguistica computazionale rappresenta uno strumento potente ma non onnipotente per investigare la preistoria linguistica. I metodi bayesiani moderni offrono rigore statistico e capacità di quantificare l'incertezza che rappresentano progressi genuini rispetto agli approcci qualitativi tradizionali. Tuttavia, soffrono di limitazioni intrinseche dovute alle assunzioni semplificatrici dei modelli, all'incompletezza e agli errori nei dati, e all'erosione inesorabile del segnale filogenetico a grandi profondità temporali.

La prospettiva più promettente non è quella di vedere la filogenesi computazionale come un sostituto dei metodi tradizionali della linguistica storica, né come la soluzione definitiva ai problemi della preistoria linguistica, ma come un componente di un approccio pluralistico che integra metodi diversi e fonti di evidenza multiple. Il metodo comparativo tradizionale, con la sua attenzione alle corrispondenze fonetiche sistematiche e alla ricostruzione dettagliata di protoforme, rimane insostituibile per stabilire relazioni genealogiche con alta confidenza. I metodi filogenetici computazionali aggiungono capacità di quantificazione, di gestione di dataset grandi, e di modellazione esplicita dell'incertezza. L'integrazione con evidenze da archeologia, genetica delle popolazioni, paleoclimatologia, e altre discipline fornisce triangolazioni che rafforzano notevolmente le conclusioni quando convergono, e segnalano problemi quando divergono. Nessuna singola linea di evidenza può risolvere definitivamente questioni complesse sulla preistoria umana, ma la convergenza di evidenze indipendenti può costruire casi cumulativi molto robusti. La ricerca futura beneficerà da miglioramenti su molteplici fronti: dataset più grandi e più accurati attraverso progetti di documentazione linguistica su larga scala, metodi computazionali più sofisticati che catturano meglio la complessità del cambiamento linguistico, calibrazioni più precise basate su cronologie archeologiche e storiche affinate, e integrazione più stretta con altre discipline. Tuttavia, è importante mantenere aspettative realistiche: esistono limiti fondamentali a quanto possiamo ricostruire della preistoria più profonda, e zone di incertezza irriducibile dovranno essere accettate piuttosto che negate. L'attitudine scientifica appropriata combina ambizione intellettuale nell'uso di metodi sofisticati per estrarre il massimo di informazione possibile dai dati disponibili, con umiltà epistemologica nel riconoscere i limiti di ciò che possiamo conoscere e nell'essere espliciti sulle incertezze e sulle assunzioni che sottendono le nostre conclusioni. In questo spirito, la filogenesi linguistica computazionale continuerà a contribuire in modo significativo alla nostra comprensione della straordinaria odissea evolutiva delle lingue umane.